The Role of Loss Functions in Regression Problems

Inaugural Dissertation
of the Faculty of Science

University of Bern

Presented by

Anja Mühlemann

from Seeberg (CH)

Supervisor of the doctoral thesis:

Prof. Dr. Johanna F. Ziegel

Institute of Mathematical Statistics and Actuarial Science
of the University of Bern

The Role of Loss Functions in Regression Problems

Inaugural Dissertation
of the Faculty of Science

University of Bern

Presented by

Anja Mühlemann

from Seeberg (CH)

Supervisor of the doctoral thesis:

Prof. Dr. Johanna F. Ziegel

Institute of Mathematical Statistics and Actuarial Science
of the University of Bern

Accepted by the Faculty of Science

Bern, 26.03.2021

The dean:
Prof. Dr. Zoltan Balogh

Acknowledgements

Throughout my PhD, I received a great deal of support from many different people.

First of all, I would like to thank my supervisor, Johanna F. Ziegel, for her precious advice and her valuable ideas throughout my PhD. I had the great opportunity to collaborate with her and thereby develop my skills as a statistician. Regardless of the nature of my questions and problems - be it related to research or be it related to my personal life - she assisted me with helpful advice and was always patient. Thank you for your support, motivation and also for your encouragement in times of doubt!

Furthermore, I would like to thank Tilmann Gneiting for the opportunity to spend six weeks at the Heidelberg Institute for Theoretical Studies and for the warm welcome. I also enjoyed collaborating with Tilmann during the current coronavirus outbreak. Thank you, for accepting to be the external referee for this thesis and your advice during our collaborations!

Besides Johanna and Tilmann, I would like to thank my coauthor Alexander I. Jordan. It was a great joy and inspiring experience to work with you. My thanks also go to Ilya Molchanov, with whom I had the pleasure to finish a research project that started with my master thesis. Thank you both for your time and motivation for our joint projects!

During the years I spent being a member of the Institute of Mathematical Statistics and Actuarial Science at the University of Bern, I had the pleasure to meet many friendly and inspiring people. Many friendships developed from these associations. In particular, I would like to thank my office mate Alexandre Mösching for his support, encouragement and friendship throughout these four years. I would also like to express my deepest appreciation for Tobias Fissler, my previous office mate, for his comments on the introduction of this thesis.

Finally, I would like to thank my family and friends, who accompanied me on this journey and spent many wonderful moments with me. In particular, I would like to thank my sister, Sarah, for her linguistic comments on some parts of this thesis and making sure that I always had enough to eat during the last few weeks. Moreover, I am enormously thankful to my boyfriend, Lukas, for his constant support and love.

Bern, March 1, 2021 *Anja Mühlemann*

Abstract

In regression analysis, the goal is to capture the influence of one or more explanatory variables $X_1, \dots, X_m$ on a response variable Y in terms of a regression function $g : \mathbb{R}^m \to \mathbb{R}$. An estimate $\hat{g}$ of g is then found or evaluated in terms of its ability to predict a prespecified statistical functional T of the conditional distribution $\mathcal{L}(Y|X_1, \dots, X_m)$. This is done with the help of a loss function that penalizes estimates that perform poorly in predicting $T(\mathcal{L}(Y|X_1, \dots, X_m))$. More precisely, it is done by using loss functions that are *consistent* for T.

Clearly, the outcome of the evaluation or estimation strongly depends on the functional T. However, when we focus on a specific functional T a vast collection of suitable loss functions may be available and the result can still be sensitive to the choice of loss function.

There are several viable solution strategies to approach this issue. We can, for instance, impose additional properties on the loss function or the resulting estimate so that only one of the possible loss functions remains reasonable. In this doctoral thesis we adopt another approach. The underlying idea is that we would naturally prefer an estimate $\hat{g}$ that is optimal with respect to several consistent loss functions for T, as then the choice of loss function seems to impact the outcome less severely.

In Chapter 1, we consider the nonparametric isotonic regression problem. We show that this regression problem is special in that for identifiable functionals T, solutions which are simultaneously optimal with respect to an entire class of consistent losses exist and can be characterized. There are, however, several functionals of interest that are not identifiable. The expected shortfall is just one prominent example. However, some of those functionals can be obtained as a function of a vector-valued elicitable functional. In the second Chapter, we investigate when simultaneous optimality with respect to a class of consistent losses holds for these functionals and introduce the solution to the isotonic regression problem for a specific loss in the case where simultaneous optimality is not fulfilled.

In parametric regression, on the other hand, different consistent loss functions often yield different parameter estimates under misspecification. This motivates to consider the set of these parameters as a way to measure misspecification. We introduce this approach in Chapter 3 and show how the set of these model parameters can be calculated on the population and on the sample level for an isotonic regression function g.

Contents

Introduction

Since the beginning of mankind, humans strive to understand and anticipate natural phenomena. While other species excel in terms of their physical strength or sharpness of senses, the ability of humans to anticipate is a quality with enormous advantages. While in the beginning, we limited ourselves to observe and generalize simple phenomena, the relationships of interest became increasingly complicated over time, so that a framework had to be created to study these relationships.

Statistics, in the modern sense of the word, has its roots in the late 19th and early 20th century. Sir Francois Galton and Karl Pearson transformed statistics - which was previously understood to be systematic collection of data - into a rigorous mathematical discipline suitable for analysis. One of Pearson's best-known contributions to the field is the Pearson correlation coefficient. Francois Galton, on the other hand, is credited for contributing key concepts such as standard deviation and even regression analysis. Later on, many other brilliant minds such as Sir Ronald A. Fisher, and William S. Gosset — just to name a few — enriched the field of statistics with notable research so that those early statistical ideas and findings remain fundamental to modern-day statistics.

This doctoral thesis primarily focuses on one of these many early statistical methods that was expanded over the years and remains vibrant to this day — regression models.

Regression models and their downsides

The goal of regression analysis is to estimate the relationship between a response variable Y and a collection of explanatory variables $X_1, \ldots, X_m$, the covariates. It is assumed that the outcome of the response depends on the outcome of the covariates so that understanding their relationship allows prediction of Y based on the outcomes of $X_1, \ldots, X_m$. For instance, we observe that taller people tend to be heavier. Understanding this relationship enables us to roughly estimate a person's weight y based on their height x_1. For simplicity, we let the response and covariates be real-valued for the course of this introduction.

Before we can predict the response from the covariates, however, the relationship between the response and the covariates has to be estimated. Of course, we desire an estimate that manages to grasp the true connection between Y and the explanatory variables $X_1, \ldots, X_m$. Thus, we strive after the *best* estimate for this relationship. What the *best* estimate is, however, depends on the context. Figure 1 contains four possible fits to a sample of 90 data points. In a situation where underestimation of the response leads to severe consequences, the dark blue line would be preferred to the others. In another context, however, the aforementioned choice may not be reasonable.

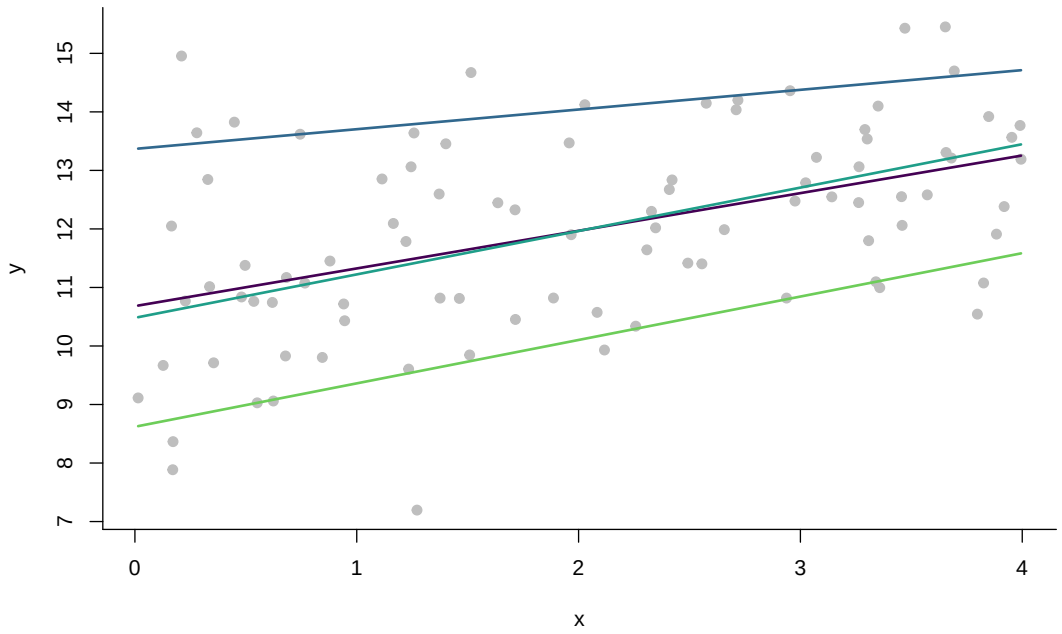

Figure 1: For a sample of 90 data points four different linear regression model fits are drawn.

Mathematical setup. In the decision-theoretic framework, we think of the word *best* in terms of a statistical functional T. Let the interval $I \subseteq \mathbb{R}$ and let $\mathcal{P}$ be a class of probability distributions on I, $P \in \mathcal{P}$. Then, a functional T is a mapping $P \mapsto T(P) \in I$. Although we could allow for set-valued functionals T, we assume that T is point-valued for this introduction to ease the exposition. Set-valued functionals can be reduced to point-valued functionals by considering the lower bound, for example. In the aforementioned context where underestimation has severe consequences, for instance, we are interested in T being an α-quantile with quantile level α close to one.

When the functional of interest has been agreed upon, the aim is to model the functional T of the conditional predictive distribution, namely

$$T(Y \mid X_1 = x_1, \ldots, X_m = x_m) = g(x_1, \ldots, x_m),$$

where $T(Y \mid X_1 = x_1, \ldots, X_m = x_m) = T(\mathcal{L}(Y|X_1 = x_1, \ldots, X_m = x_m))$ slightly abusing notation for the sake of brevity. The function $g : \mathbb{R}^m \to \mathbb{R}$ is called the *regression function*. In quantile regression, one seeks to model the conditional quantile of the response given the covariates. In ordinary least squares regression, on the other hand, the goal is to model the conditional mean instead.

To model $T(Y \mid X_1 = x_1, \ldots, X_m = x_m)$ one uses training data

$$\{(x_{i1}, \ldots, x_{im}, y_i) : i = 1, \ldots n\}$$

to estimate a parametric or non-parametric statistical model $\hat{g}$ for g. Then, one uses $\hat{g}(x_1, \ldots, x_m)$ as a point estimate for a future realization of the response variable Y, given the specific realizations $x_1, \ldots, x_m$ of the explanatory variables at hand.

Evaluation of an estimate. To assess whether the estimate $\hat{g}$ is a good fit for g an error measure is required. In practice, estimates are determined or compared by means of a loss function

$$L : I \times I \to [0, \infty)$$

that depends on both, the estimate and the response. We take the loss function to be *negatively oriented* which means the smaller the loss, the better the estimate. Prime examples of loss functions include the squared error and the absolute error.

Clearly, the choice of the loss function should reflect the choice of the functional T, in that estimates that are better at modeling T should yield a smaller loss. In other words, we are looking for a loss function L that is *consistent* for functional T. Formally, we say that a loss function L is consistent for functional T if

$$\mathbb{E}L(T(P), Y) \leq \mathbb{E}L(\hat{y}, Y)$$

for all P, and all $\hat{y} \in I$, where Y is a random variable with distribution P. It is strictly consistent if it is consistent and equality implies that $\hat{y} = T(P)$. Following Osband (1985) and Lambert et al. (2008), a functional is *elicitable* if there exists a loss function that is strictly consistent for it.

However, there is usually a wide range of consistent loss functions for functional T available. The following example gives an insight into the variety of consistent loss function for the mean and the α-quantile.

Example. (a) Savage (1971) showed that under mild regularity conditions the entire class $\mathcal{L}$ of Bregman losses, that is, all loss functions of the type

$$L_\phi(\hat{y}, y) = \phi(y) - \phi(\hat{y}) - \phi'(\hat{y})(y - \hat{y}),$$

where ϕ is a convex function with subgradient ϕ', is consistent for the mean functional $\mathbb{E}$. Its most prominent member is the squared error. Subject to weak regularity conditions the class of Bregmann losses actually comprises all consistent loss functions for the mean.

(b) The class of consistent loss functions for quantiles has been characterized by Thomson (1979) and Saerens (2000) and is given by all loss functions of the form

$$L(\hat{y}, y) = (\mathbb{1}\{y \leq \hat{y}\} - \alpha)(h(\hat{y}) - h(y)),$$

where $h : I \to \mathbb{R}$ is a nondecreasing function, $\alpha \in (0, 1)$.

A characterization of the class of consistent losses for expectiles and a comprehensive and more detailed discussion on the topic of issuing and evaluating point forecasts can be found in Gneiting (2011).

Choice of loss function. While the choice of the functional T certainly reduces the number of suitable losses, we may still choose from a substantially large class. Typically, there is no argument to prefer one consistent loss over the others in terms of accuracy. Only imposing additional restrictions may lead to a clear choice. For T being the mean functional, the Gauss-Markov Theorem suggests that from an efficiency standpoint the quadratic loss would be preferable under homoscedasticity. The quadratic loss is also the only Bregmann loss function that solely depends on the difference of the target y and the estimate $\hat{y}$ (Savage, 1971). Thus, if this property is desired then the squared error should be chosen. The QLIKE loss

$$L(\hat{y}, y) = \frac{y}{\hat{y}} - \log \frac{y}{\hat{y}} - 1,$$

on the other hand, is the only Bregman loss function that solely depends on the ratio of the target y and the estimate $\hat{y}$ (Patton, 2011).

To elaborate on the impact of the choice of loss function, let us quickly recapitulate the standard approach to parametric regression. To estimate g, one selects a suitable parametric forecasting model,

$$m : \mathbb{R}^m \times \Theta \to \mathbb{R}, \quad (x_1, \ldots, x_m, \theta) \mapsto m\left(x_1, \ldots, x_m; \theta\right), \tag{1}$$

where Θ is the set of admissible parameters. Hereafter, one minimizes the expected loss of some specific loss function L on the training data to obtain an estimate $\hat{\theta}_n(L)$ of the model parameter, that is,

$$\hat{\theta}_n(L) := \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\Big(m(x_{i1}, \ldots, x_{im}; \theta), y_i\Big). \tag{2}$$

If the model is correctly specified for functional T, that is,

$$T\left(Y|X_1, \ldots, X_m\right) = m\left(X_1, \ldots, X_m; \theta^*\right) \quad \text{almost surely for some } \theta^* \in \mathbb{R}^p,$$

minimization of the above criterion yields a consistent estimator of θ^* for *any* choice of consistent loss function, subject to moment conditions and the assumption that the data are independent and identically distributed. In finite samples, the estimators usually differ, and under model misspecification, even their limits may differ. Therefore, the model parameter estimate is sensitive to the choice of the loss function used in estimation.

For T being the mean functional Patton (2020) demonstrates that also the ranking of forecasts is sensitive to the choice of consistent loss function. This sensitivity only vanishes in the absence of model misspecification and estimation error. But in almost all practical applications one of these complications occurs. Even worse, models are usually simplifications and can thus not capture the true relationship in its entirety. This misspecification may not necessarily lead to unlikely forecasts but it can.

In ordinary least squares regression, for example, the estimates are traditionally obtained by minimizing the squared error loss. However, any Bregman loss would be a reasonable choice for the loss function L. Nevertheless, estimation by minimizing (2) for a Bregman loss function different from the quadratic loss has rarely been attempted. Thus, we may ask ourselves how different our interpretations would have turned out when another Bregman loss would have been chosen instead. Figure 2 contains three simple linear regression fits, each obtained by minimizing a different Bregman loss. While, for this sample, in the correctly specified case the differences are not yet substantial, this ceases to be true when the underlying relationship is no longer linear.

Possible solution strategies

It is of course unpleasant to see that the choice of loss L may impact our conclusion so drastically. Possible solution strategies involve defining additional desiderata on the estimator or loss, so that the choice for suitable loss function is clear. One of them is certainly the efficiency of the estimate. Another possible strategy is to choose an estimate that is optimal with respect to several consistent loss functions. That way the impact of the decision made appears to be less severe. For this strategy, estimates that are

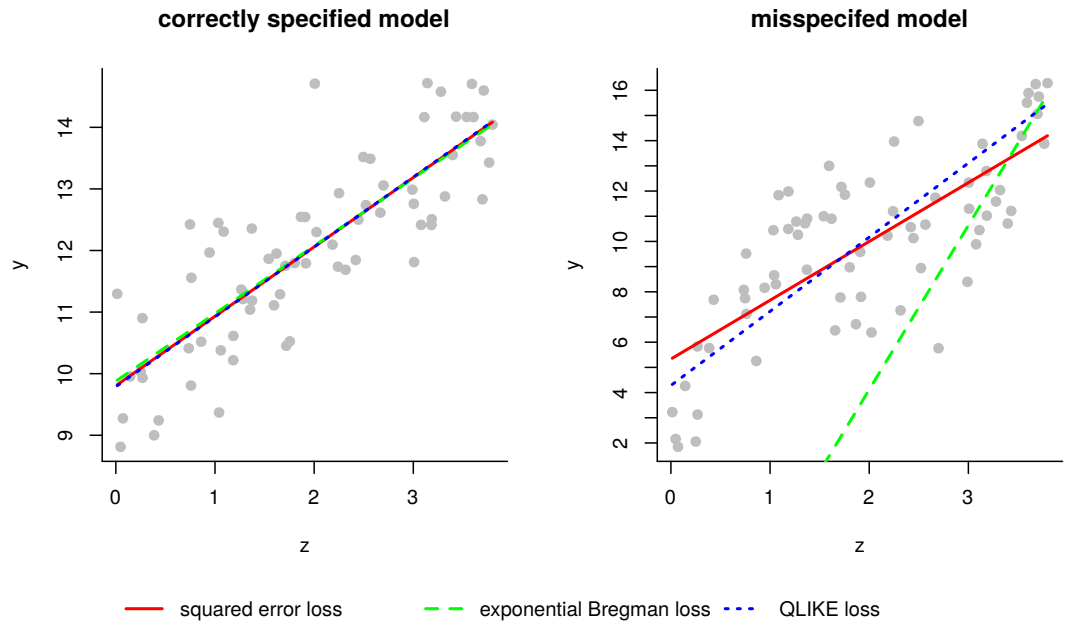

Figure 2: On the left, a sample of 70 data points generated from a linear relationship is drawn. On the right, the underlying relationship of the data points is cubic. For both samples the linear regression fits obtained by minimizing the squared error loss (red), the QLIKE loss (blue, dotted) and the exponential Bregman loss $L(\hat{y}, y) = \exp(y) - \exp(\hat{y}) - \exp(\hat{y})(y - \hat{y})$ (green, dashed) are drawn.

simultaneously optimal for an entire class $\mathcal{L}$ of loss functions are desirable to eliminate or at least reduce the impact of the choice in the results.

Simultaneous optimality for identifiable functionals. Sometimes not enough is known about g to choose a specific statistical model in the spirit of (1). Nonetheless, some properties of the relationship may be known. In this case, we could resort to nonparametric regression under shape constraints. In the following, we consider the special case of only one covariate. It could be known that the relationship is increasing in that larger values of the covariate X result in larger values of $T(Y|X)$. But since we are unaware whether the relationship is linear, quadratic, etc., we cannot settle on a specific model. The only constraint on $\hat{g}$ is that it is increasing. For this specific shape constraint, called *isotonicity*, Barlow et al. (1972) showed that there exists a solution $\hat{g} : \{x_1, \ldots, x_n\} \mapsto \mathbb{R}$ that is simultaneously optimal for all Bregman losses L_ϕ in that it minimizes the expected loss for all $L_\phi \in \mathcal{L}$. This means that if the functional we agreed upon is the mean functional $\mathbb{E}$ then a solution optimal for all consistent loss functions exists. This solution can efficiently be calculated by the pool-adjacent violators (PAV) algorithm and was developed independently by several parties in the 1960s (Ayer et al., 1955; Bartholomew, 1959a,b; Brunk, 1955; van Eeden, 1958; Miles,

1959). An extensive survey of the history of the PAV-algorithm can be found in de Leeuw et al. (2009). Brümmer and Du Preez (2013) rediscover the result of Barlow et al. (1972) that the PAV-algorithm leads to a simultaneously optimal solution for all proper scoring rules in the context of binary events. Later on, this result was generalized to quantile functionals T in Robertson and Wright (1973) and Robertson and Wright (1980), yet they treat quantiles as point-valued functionals. Recently, Henzi et al. (2019) derived a nonparametric method to estimate the conditional distribution under the isotonic shape constraint that is simultaneously optimal relative to a class of relevant loss functions. Because stochastic orders can equivalently be defined in terms of quantiles, they obtain isotonic quantile regression for free. When talking about stochastic order and its equivalent formulation in terms of quantiles, it is worth mentioning the paper of Mösching and Dümbgen (2020) who investigate the two approaches and derive min-max and max-min formulas as lower and upper bounds for the optimal isotonic solution in the context of set-valued minimizers of convex and coercive loss functions.

A disadvantage of the aforementioned estimates is that they are only defined on the observed values of the covariate X. Of course, $\hat{g}$ can be extended to the entire covariate space by interpolation. Unfortunately, the resulting function is rather rough. Mammen (1991) shows that applying smoothing before or after the PAV-algorithm leads to asymptotically equivalent results. However, performing smoothing before applying the PAV-algorithm leads to a smaller quadratic loss if the kernel function is not too smooth.

In the first chapter of this thesis, we complement the existing work by a complete characterization of all simultaneously optimal solutions for any functional that can be defined via an identification function. Thereby, we recover aforementioned results for the mean functional and quantile functional. Our approach is based on the possibility to write any consistent loss function as a mixture of elementary scores. Ehm et al. (2016) introduced this result for quantiles and expectiles. However, the result can be extended to any functional T that is identifiable with an oriented identification function; see Gneiting (2011); Steinwart et al. (2014); Ziegel (2016b). But contrary to quantiles and expectiles the class of losses admitting a mixture representation may not comprise the entire class of consistent loss functions for T. Nonetheless, it still encompasses a considerable amount of different loss functions. We additionally generalize our results to partially ordered explanatory variables. Moreover, we note that under weaker shape constraints simultaneous optimality may be unattainable even if the shape constraint of isotonicity is only slightly weakened as it is the case for unimodal regression.

The results in Chapter 1 only apply functionals T that can be defined via an identification function. Steinwart et al. (2014) showed that under certain regularity assumptions identifiability and elicitability are equivalent for univariate functionals. Therefore identifiability as a restriction is not too severe. An exception to this equivalence that is sometimes of interest is the mode functional.

However, there are several functionals of interest are not elicitable and thus certainly not identifiable. For instance, the expected shortfall risk measure is appealing for application in financial contexts as it is coherent, yet not elicitable.

Simultaneous optimality for Bayes risks. As mentioned, our results in the first chapter of this thesis do not apply to nonelicitable functionals. Fortunately, there is a solution to this. Given an identifiable functional, Osband (1985) equips us with a tool known as *Osband's principle* to describe and characterize the class of the loss functions that are consistent for said functional. Fissler and Ziegel (2016) refine and generalize Osband's principle which enables them to characterize class of all strictly consistent loss functions for vectors of quantiles or (and) expectiles at different levels. Furthermore, they show that the expected shortfall is jointly elicitable with the value-at-risk, the quantile, and characterized the corresponding class of all strictly consistent loss functions. Frongillo and Kash (2020) extend this result. They show that given a fixed (possibly multivariate) elicitable functional T and any strictly consistent loss L for T, its Bayes risk, that is, the minimizer of the expected loss, is jointly elicitable with T. Furthermore, they characterize a corresponding class of consistent loss functions. This class may, however, be a strict subset of the class of consistent loss functions. For the pair value-at-risk and expected shortfall they recover the result of Fissler and Ziegel (2016). Nolde and Ziegel (2017) describe the classes of strictly consistent and homogeneous loss functions for the pair value-at-risk and expected shortfall and Ziegel et al. (2020) introduce a Choquet-type mixture representation analogous to Ehm et al. (2016) for the pair.

In the second chapter of this thesis, we aim to combine and extend these results to see whether simultaneous optimality results, analogous to Chapter 1, also hold for functionals with *elicitation complexity* greater than one, that is, for functionals that are not elicitable themselves but the Bayes risk of an elicitable functional. To this end, we describe the solutions to this isotonic regression problem and study whether these solutions are simultaneously optimal. Unfortunately, simultaneous optimality is not necessarily attainable for these functionals. Even when considering vectors with elicitable components simultaneous optimality may only hold for a strict subset of the class of consistent loss functions. An exception to this phenomenon is given by vectors of quantiles where simultaneous optimality with respect to the entire class of consistent losses continues to hold. This explains some of the optimality results of Henzi et al. (2019). Nevertheless, we derive a criterion to determine whether a solution at hand is indeed simultaneously optimal. With a simulation study, we investigate how often the solution obtained is not simultaneously optimal. Sadly, this occurs rather frequently. Motivated by this discovery, we additionally dedicate our attention to optimal isotonic solutions for a specific loss function. In Chapter 3, we put our method into practice. To this end, we aim to predict the value-at-risk and the expected shortfall of the negated log-returns of the NASDAQ Composite Index. It is assumed that the corresponding volatility index has an isotonic relationship with the negated log-returns so that we use

today's volatility to forecast tomorrow's value-at-risk and expected shortfall. We compare our forecasts to forecasts obtained by the methods used by Nolde and Ziegel (2017) and Patton et al. (2019) and discuss possible improvements. Even if the forecasts obtained via isotonic regression cannot fully compete with the others, we are amazed by how well our simple method performs.

Pareto-optimal parameters. So far, we only considered nonparametric models. But what can be done if we are interested in a parametric model instead? We have previously seen that models are often simplifications of reality. In some cases the predictions remain reliable, while in others the misspecification impacts the predictions. Often, it is impossible to prevent model misspecification (Patton, 2011). Therefore, it is fundamental to develop methods that provide reasonable results or inform us about the reliability of our conclusions, even in the presence of misspecification.

Point estimates tend to convey a false sense of security. Thus, many agree that forecasts should be of probabilistic nature, in other words, they should quantify their own uncertainty additionally to stating the predicted outcome. But sometimes we are interested in point estimates instead. In such cases confidence intervals are disclosed to reflect the uncertainty. Hansen et al. (2011) adopt this strategy to introduce a model confidence set that contains the best model with a given confidence. Sometimes a portion of the uncertainty is due to decisions yet to be made. In such cases, Feiler and Ajdler (2019) suggest to include the relations among competing models into the considerations to reduce uncertainty. Another possible approach, pursued by Holland (2019) for mean estimation, is to develop methods with fewer assumptions so that misspecification is less likely to occur. In Bayesian statistics, the task of developing methods more robust to misspecification has gained a lot of attention recently. Huggins and Miller (2019) use bootstrap to obtain more robust posteriors, Thomas and Corander (2019), on the other hand, use tempering instead, and Loaiza-Maya et al. (2019) replace the standard Bayesian up-date by a criterion that captures a user-specified measure of predictive accuracy. However, Bayesian inference is by far not the only approach sensitive to model misspecification. Buja et al. (2019) examine the consequences of nonlinearity when considering linear models. They show that in the presence of nonlinearity the covariates can no longer be treated as fixed. The randomness of the covariates, however, does affect the parameter estimates and creates sampling variability.

In the third chapter of this thesis, we introduce a novel approach to capture model misspecification. Our approach is closely related to the observations made by Buja et al. (2019). Considering simple linear regression, we introduce the set of *Pareto-optimal* model parameters, that is, the set of all parameters that are not *strictly dominated* by another parameter, where a parameter θ_1 is said to be strictly dominated by a parameter

θ_2, relative to class $\mathcal{L}$, if

$$\mathbb{E}L\left(m(X;\theta_2),Y\right) \leq \mathbb{E}L\left(m(X;\theta_1),Y\right), \quad \text{for all } L \in \mathcal{L},$$

where the above inequality is strict for at least one $L \in \mathcal{L}$. We show how the set of Pareto-optimal parameters increases in size under misspecification and it can explicitly be calculated on the population level in the case of isotonic regression. Interestingly, on the population level, the linear models given by the Pareto-optimal parameters correspond to the tangents and chords of g. This is exactly what Buja et al. (2019) observed concerning the sampling variability. Moreover, relying on the results in Chapter 1, we also succeeded in calculating the Pareto-optimal set on the sample level. Finally, we put our methodology to the test and assess model misspecification in two data examples.

1 Optimal solutions to the isotonic regression problem

Alexander I. Jordan, Anja Mühlemann and

Johanna F. Ziegel

Abstract. In general, the solution to a regression problem is the minimizer of a given loss criterion, and depends on the specified loss function. The nonparametric isotonic regression problem is special, in that optimal solutions can be found by solely specifying a functional. These solutions will then be minimizers under all loss functions simultaneously as long as the loss functions have the requested functional as the Bayes act. For the functional, the only requirement is that it can be defined via an identification function, with examples including the expectation, quantile, and expectile functionals.

Generalizing classical results, we characterize the optimal solutions to the isotonic regression problem for such functionals, and extend the results from the case of totally ordered explanatory variables to partial orders. For total orders, we show that any solution resulting from the pool-adjacent-violators algorithm is optimal. It is noteworthy, that simultaneous optimality is unattainable in the unimodal regression problem, despite its close connection.

Acknowledgments. We would like to thank Tilmann Gneiting, Alexandre Mösching and Lutz Dümbgen for inspiring discussions and valuable comments. Anja Mühlemann and Johanna F. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation.

1.1 Introduction

Suppose that we have pairs of observations $(z_1, y_1), \ldots, (z_n, y_n)$ where we assume that y_i, $i = 1, \ldots, n$ are real-valued. The aim of isotonic regression is to fit an increasing function $\hat{g}\colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ to these observations. The covariates $z_1, \ldots, z_n$ can take values in any set as long as it is equipped with a partial order which we denote by $\preceq$. Then, a function $g\colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ is *increasing* if $z_i \preceq z_j$ implies that $g(z_i) \leq g(z_j)$.

As it is common in regression analysis, we aim to find an estimate $\hat{g}$ that minimizes the expected loss for some loss function $L\colon \mathbb{R} \times \mathbb{R} \to [0, \infty)$. If the function $\hat{g}$ is interpreted as an estimator of the conditional expectation of a random variable Y given Z, then a natural choice for L is the squared error loss $L(x, y) = (x - y)^2$. For $i \leq j$, let $\mathbb{E}_{i:j}$ denote the expectation with respect to the empirical distribution of $(z_i, y_i), \ldots, (z_j, y_j)$. Assuming that $z_1 < z_2 < \cdots < z_n$, the minimizer of the quadratic loss criterion

$$\mathbb{E}_{1:n}(g(Z) - Y)^2 \tag{1.1}$$

over all increasing functions g is given by

$$\hat{g}(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} \mathbb{E}_{i:j} Y = \max_{i \leq \ell} \min_{j \geq i} \mathbb{E}_{i:j} Y, \quad \ell = 1, \ldots, n, \tag{1.2}$$

see Barlow et al. (1972, eq. (1.9)–(1.13)). The solution $\hat{g}$ can be computed efficiently using the so-called pool-adjacent-violators (PAV) algorithm. These results were developed in the 1950s by several parties independently; see Ayer et al. (1955), Bartholomew (1959a), Bartholomew (1959b), Brunk (1955), van Eeden (1958), Miles (1959).

It turns out that the solution given at (1.2) is also the unique minimizer of the Bregman loss criterion

$$\mathbb{E}_{1:n} L(g(Z), Y), \tag{1.3}$$

where the squared error loss in (1.1) has been replaced by a Bregman loss function $L = L_\phi$ (Barlow et al., 1972, Theorem 1.10). That is,

$$L_\phi(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x),$$

where ϕ is a convex function with subgradient ϕ'. Savage (1971) found that the Bregman class comprises all loss functions L where the expectation functional minimizes the expected loss, i.e.,

$$\mathbb{E}_P Y = \arg\min_x \mathbb{E}_P L(x, Y),$$

where Y is a random variable with distribution P. Due to this property, any loss function in the Bregman class is also referred to as a consistent loss function for the expectation functional (Gneiting, 2011).

In summary, the increasing regression function at (1.2) is simultaneously optimal with respect to all consistent loss functions for the expectation. This robustness with respect to the choice of loss function means that the solution to the regression problem is determined by the choice of the expectation as the target functional. We will see that the same holds for other functionals. As such, in nonparametric isotonic regression we can replace the task of choosing a loss function with the task of choosing a suitable target functional.

This remarkable result is particularly beneficial in scenarios where a single relevant loss function cannot easily be identified. For example, institutions such as central banks or weather services provide analyses and forecasts that drive individual decision making in a heterogeneous group of users. In these circumstances, determining a unifying loss function is hardly trivial. However, publishing results for the expectation and for various quantile levels is certainly feasible.

The simultaneous-optimality result for nonparametric isotonic regression is in stark contrast to the optimality behavior of parametric models for increasing regression functions. Suppose that $\{g_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$ is a parametric model of increasing functions g_θ. Then, the optimal parameters with respect to the Bregman-loss criterion (1.3) generally vary (substantially) depending on the chosen loss function (Patton, 2020). Consistency of the loss function merely ensures that the true parameter value of a correctly specified model minimizes the Bregman-loss criterion on the population level. Interestingly, simultaneous optimality with respect to all consistent loss functions generally also breaks down if one weakens the isotonicity constraint of the regression function to a unimodality constraint; see Section 1.3.

In this paper, we generalize the result of Barlow et al. (1972, Theorem 1.10) in several directions. First, instead of the expectation functional, we consider general (possibly set-valued) functionals T that are given by an identification function $V(x, y)$ as defined in Definition 1.2.1. Second, in the case of set-valued functionals, we give a complete characterization of all possible solutions. Third, we demonstrate that a suitably modified version of min-max or max-min solutions as in (1.2) continues to hold for general partial orders on the covariates.

An identification function is an increasing function that weighs negative values in the case of underestimation against positive values in the case of overestimation, with an optimal expected value of zero. The corresponding functional T then maps to the optimizing argument (or set of optimizing arguments). Prime examples of such functionals are (possibly set-valued) quantiles, expectiles (Newey and Powell, 1987), or ratios of expectations. Quantiles, including the median, have previously also been treated in Robertson and Wright (1973, 1980), but not in the interpretation as set-valued functionals. Predefining a global scheme for reducing the median interval to a single point (e.g., some weighted average of lower and upper functional value) inevitably restricts the possible solutions to the isotonic regression problem. Figure 1.1 illustrates this issue, and shows

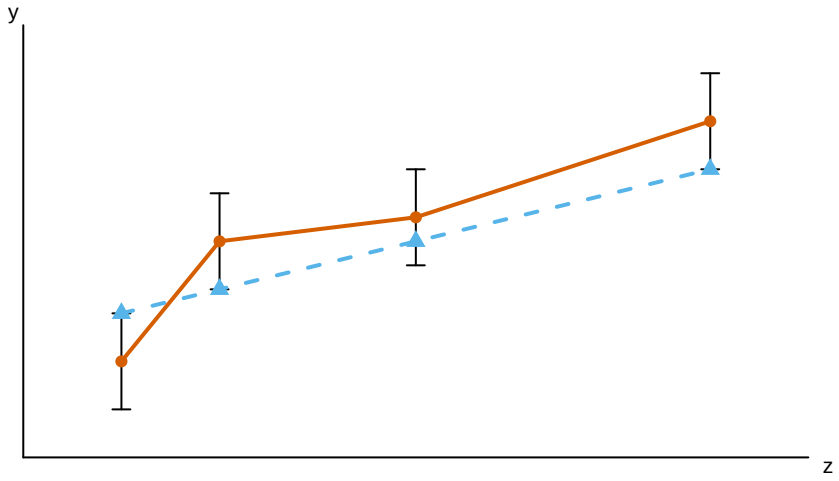

Figure 1.1: **Solutions in isotonic quantile regression.** Two solutions to the isotonic regression problem are shown for an example with $z = z_1, \ldots, z_4$. The red curve is predetermined to pass through the midpoint of the functional intervals, whereas the blue curve illustrates the smoothest solution with minimal slope.

how a more general interpretation of the functional as set-valued facilitates solutions with secondary optimality criteria such as smoothness and minimal slope. Expectiles and ratios of expectations, on the other hand, have been fully treated in Robertson and Wright (1980). These functionals map to single values and satisfy the Cauchy mean value property which is implied by identifiability.

In contrast to previous work, we treat all functionals as set-valued. In Section 1.4, we give explicit solutions for the lower and upper bound of the isotonic regression problem in the context of partial orders. The method of proof for these results is fundamentally different from the approach of Barlow et al. (1972, Theorem 1.10) or Robertson and Wright (1980), and in contrast to the latter comes with an immediate construction principle for loss functions. Our method relies on the mixture or Choquet representations of consistent loss functions, introduced by Ehm et al. (2016) for the quantile and expectile functionals. Given the identification function $V(x, y)$ for the functional T, a one-parameter family of elementary loss functions that are consistent for the functional T can be readily defined,

$$S_\eta(x, y) = (\mathbb{1}\{\eta \leq x\} - \mathbb{1}\{\eta \leq y\})\, V(\eta, y),$$

where $\eta \in \mathbb{R}$. For all consistent loss functions L in the class

$$\mathcal{S} = \left\{ \int_{\mathbb{R}} S_\eta(x, y)\, \mathrm{d}H(\eta) : H \text{ is a nonnegative measure on } \mathbb{R} \right\}, \tag{1.4}$$

the optimal isotonic solution to the criterion (1.3) is bounded below by a min-max formula and bounded above by a max-min formula as in (1.2) with the expectation

replaced by the lower and upper functional values under T, respectively. We show that the min-max or max-min solution is simultaneously optimal with respect to all elementary loss functions for T, and hence with respect to the entire class $\mathcal{S}$. In fact, optimality of an isotonic solution with respect to the criterion (1.3) for $L = S_\eta$ for some $\eta \in \mathbb{R}$ corresponds to finding a solution with optimal superlevel set $\{g \geq \eta\}$. Considering an isotonicity constraint as a constraint on admissible superlevel sets of the regression function relates to the work of Polonik (1998) in the context of density estimation.

If T is a quantile, an expectile, or a ratio of expectations, then $\mathcal{S}$ comprises all consistent loss functions for T subject to standard conditions, and if $V(x, y) = x - y$ is the identification function of the expectation, then the class $\mathcal{S}$ is the class of Bregman loss functions; see Ehm et al. (2016); Gneiting (2011). We also give results that can be directly translated to a simple algorithm that recovers the full range of optimal solutions from the lower and upper bounds and the full data set. While the bounds alone do not contain sufficient information, only few additional computations on the entire data set are necessary. Our method of proof also leads to a transparent proof of the validity of the PAV algorithm; see Section 1.4.2.

Recently, Mösching and Dümbgen (2020) derived a similar result of min-max and max-min formulas as lower and upper bounds for optimal isotonic solutions in the context of set-valued minimizers of convex and coercive loss functions. Brümmer and Du Preez (2013) rediscover the result of Barlow et al. (1972) that the PAV algorithm leads to a simultaneously optimal solution for all proper scoring rules in the context of binary events – a special class of loss functions that are consistent for the expectation functional.

In Section 1.4, we treat general partial orders on the covariates and demonstrate that a suitably modified version of min-max or max-min solutions continues to hold. Again, the optimal isotonic fit is simultaneously optimal with respect to all loss functions in $\mathcal{S}$ defined at (1.4). The results in Robertson and Wright (1980) not only hold for a large class of functionals, but also for partial orders on the covariates. However, the generality of their results is limited by treating potentially set-valued functionals as maps to single values. To the best of our knowledge, the literature following Robertson and Wright (1980) is void of further results that characterize the solutions to the isotonic regression problem, or any investigations into the effect of the choice of loss function among options sharing the same Bayes act.

A comprehensive overview on isotonic regression is given in the monograph Groeneboom and Jongbloed (2014). Also, Guntuboyina and Sen (2018) review risk bounds, asymptotic theory, and algorithms in common nonparametric shape-restricted regression problems in the context of least squares optimization. Among the most recent developments on algorithms for isotonic regression with partially ordered covariates, Kyng et al. (2015) and Stout (2015) provide fast algorithms for isotone regression under different loss functions using the representation of a partial order as a directed acyclic graph. Recent advances

on asymptotic theory for isotonic regression include Han et al. (2019), giving rates for least squares isotonic regression on the unit cube of arbitrary dimension, and Bellec (2018), considering isotonic, unimodal, and convex regression in the context of total orders. Another recent interest is the regularization of isotonic regression on multiple variables with Luss and Rosset (2017) proposing a method via range restriction on the solution to the regression problem.

1.2 Functionals and consistent loss functions

We start with the definition of a functional via an identification function.

Definition 1.2.1. A function $V\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is called an *identification function* if $V(\cdot, y)$ is increasing and left-continuous for all $y\in\mathbb{R}$. Then, for any finite and nonnegative measure P on $\mathbb{R}$, we define the *functional* T induced by an identification function V as

$$T(P) = [T_P^-, T_P^+] \subseteq [-\infty, +\infty] = \bar{\mathbb{R}},$$

where the lower and upper bounds are given by

$$T_P^- = \sup\{x : V(x,P) < 0\} \quad \text{and} \quad T_P^+ = \inf\{x : V(x,P) > 0\},$$

using the notation $V(x,P) = \int_{-\infty}^{\infty} V(x,y)\,\mathrm{d}P(y)$, and assuming that all relevant integrals exist.

Defining functionals for any finite and nonnegative measure, as opposed to merely probability distributions, is a minor detail that simplifies notation when joining and intersecting data subsets. Except in the case of the null measure, any finite and nonnegative measure can be replaced with a corresponding probability distribution, without any change to the functional values.

As previously mentioned in Definition 1.2.1, we henceforth assume that the integral $V(x,P)$ exists for all relevant covariate values x. Note that T_P^- can take the value $-\infty$ and T_P^+ can take the value $+\infty$. In the subsequent results, we repeatedly refer to the smallest or largest element of a finite set where one of the elements could be $\pm\infty$. We still write min and max of the set but this quantity could be $\pm\infty$.

Definition 1.2.2. A functional T is called a *functional of singleton type* if $T(P)$ is a singleton whenever P is not the null measure. Otherwise, T is called a *functional of interval type*.

Table 1.1 summarizes common functionals and their respective identification functions, and Example 1.2.3 explains two options in more detail.

Table 1.1: **Selection of functionals and their respective identification functions.** The parameters satisfy $\alpha, \tau \in (0,1)$, $p > 1$ and $\delta > 0$, and $u : I \to \mathbb{R}$ and $w : I \to (0, \infty)$ are measurable functions on an interval $I \subseteq \mathbb{R}$. The functionals "ℓ_p minimizer" and "Huber minimizer" map to the intervals of values minimizing the ℓ_p loss and the Huber loss (Huber, 1964), respectively.

Functional	Identification function	Type
Median	$V(x,y) = \mathbb{1}\{x > y\} - 1/2$	interval
Mean	$V(x,y) = x - y$	singleton
2^{nd} Moment	$V(x,y) = x - y^2$	singleton
α-Quantile	$V(x,y) = \mathbb{1}\{x > y\} - \alpha$	interval
τ-Expectile	$V(x,y) = 2\lvert\mathbb{1}\{x > y\} - \tau\rvert(x - y)$	singleton
Ratio $\mathbb{E}_P(u(Y))/\mathbb{E}_P(w(Y))$	$V(x,y) = xw(y) - u(y)$	singleton
ℓ_p minimizer	$V(x,y) = \operatorname{sign}(x-y)\lvert x-y\rvert^{p-1}$	singleton
Huber minimizer	$V(x,y) = \operatorname{sign}(x-y)\min(\lvert x-y\rvert, \delta)$	interval

Example 1.2.3. Let $\alpha, \tau \in (0,1)$, and let P denote a probability distribution.

(a) Consider the identification function $V(x,y) = \mathbb{1}\{x > y\} - \alpha$, then $V(x,P) = P(Y < x) - \alpha$, and the interval of all α-quantiles of P,

$$T(P) = [\sup\{x : P(Y < x) < \alpha\}, \inf\{x : P(Y < x) > \alpha\}],$$

is potentially of positive length.

(b) The identification function $V(x,y) = 2\lvert\mathbb{1}\{x > y\} - \tau\rvert(x - y)$ leads to

$$V(x,P) = 2(1-\tau)\int_{-\infty}^{x}(x-y)\,\mathrm{d}P(y) + 2\tau\int_{x}^{\infty}(x-y)\,\mathrm{d}P(y),$$

which is strictly increasing and continuous in its first argument. Hence, there exists a unique solution in x for the equation $V(x,P) = 0$, and we call that solution the τ-expectile $e_\tau(P)$. In particular, for $\tau = \frac{1}{2}$ we obtain $V(x,y) = x - y$ and thus $T(P) = \{\mathbb{E}_P(Y)\}$.

In the later proofs, we use three implications of Definition 1.2.1 repeatedly to establish order relationships between the variable in the first argument of V and the functional of an empirical distribution. To facilitate reference, we note these statements explicitly.

Corollary 1.2.4. *Let V be an identification function inducing the functional T, and P be a finite and nonnegative measure on $\mathbb{R}$. Then,*

$$\begin{aligned} V(\eta, P) = 0 &\implies \eta \in T(P), \\ V(\eta, P) > 0 &\implies \eta > \sup T(P) = T_P^+, \\ V(\eta, P) < 0 &\implies \eta \leq \inf T(P) = T_P^-. \end{aligned}$$

Lemma 1.2.5 shows that a generalized version of the Cauchy mean value property, used to define functionals in Robertson and Wright (1980), holds for any functional we consider in this paper. This suggests that our results are less general, unless it can be proven that every Cauchy mean value function can be defined in terms of an identification function. On the other hand, in contrast to Robertson and Wright (1980), we treat set-valued functionals and their boundaries rigorously, and retain a higher level of generality in that regard.

Lemma 1.2.5. *Let P, Q be finite and nonnegative measures on $\mathbb{R}$. Then,*

$$\min\{T_P^-, T_Q^+\} \leq T_{P+Q}^- \leq T_{P+Q}^+ \leq \max\{T_P^-, T_Q^+\}.$$

Proof. The statement follows from Definition 1.2.1. The second inequality is trivial. For the first inequality, and $x < \min\{T_P^-, T_Q^+\}$, we have $V(x, P) < 0$ and $V(x, Q) \leq 0$, hence $V(x, P + Q) < 0$. A similar argument applies to the third inequality. □

The definition of a functional in terms of an identification function comes with a straightforward construction principle for large classes of loss functions. In a nutshell, a continuous oriented identification function defines a functional via its unique root in the first argument, a first-order condition. By integration, corresponding loss functions inherit the consistency for the functional, i.e., the minimum expected loss is attained by any member in $T(P)$. The loss functions defined in Proposition 1.2.6 are the most basic, in the sense that they are a result of integration with respect to the Dirac measure at a given threshold $\eta \in \mathbb{R}$. A similar result has also been discussed in Dawid (2016) and Ziegel (2016b).

Proposition 1.2.6. *Let V be an identification function, T be the induced functional, and $\eta \in \mathbb{R}$. Then the elementary loss function $S_\eta \colon \bar{\mathbb{R}} \times \mathbb{R} \to \mathbb{R}$ given by*

$$S_\eta(x, y) = (\mathbb{1}\{\eta \leq x\} - \mathbb{1}\{\eta \leq y\})\, V(\eta, y)$$

is consistent *for T relative to the class $\mathcal{P}$ of probability distributions with finite support. That is,*

$$\mathbb{E}_P S_\eta(t, Y) \leq \mathbb{E}_P S_\eta(x, Y), \quad \text{for all } P \in \mathcal{P},\ \text{all } t \in T(P) \text{ and all } x \in \bar{\mathbb{R}}.$$

Proof. Let

$$d(\eta) = \mathbb{E}_P S_\eta(t, Y) - \mathbb{E}_P S_\eta(x, Y) = (\mathbb{1}\{\eta \leq t\} - \mathbb{1}\{\eta \leq x\})\, V(\eta, P).$$

If $V(\eta, P) = 0$ then $d(\eta) = 0$. If $V(\eta, P) < 0$ it follows from Corollary 1.2.4 that $\eta \leq t$ and therefore $d(\eta) \leq 0$. Similary, if $V(\eta, P) > 0$ it follows that $\eta > t$ and therefore $d(\eta) \leq 0$. □

As an immediate consequence of the consistency of elementary loss functions for the functional T, we have that all loss functions in the class $\mathcal{S}$ defined at (1.4) are also consistent for the functional T. This result exemplifies an important line of reasoning used multiple times in this paper: A property of S_η that holds for all $\eta \in \mathbb{R}$ translates to the class $\mathcal{S}$.

The importance of the construction in Proposition 1.2.6 lies in the postponing of integration, or, in other words, applying Fubini in a double integration (with respect to P and to H), and then showing the property of consistency for the integrand S_η for each η rather than for the original loss function which is the integral of S_η with respect to $\mathrm{d}H(\eta)$.

Table 1.2: **Commonly used loss functions that are consistent for the mean functional.** For an interval $I \subseteq \mathbb{R}$, a Bregman loss is induced by a convex function $\phi : I \to \mathbb{R}$ with subgradient ϕ'. See Patton (2011, 2020) for the QLIKE loss and the exponential Bregman loss, respectively.

Name	Mixing measure $H((\eta_1, \eta_2]) =$	Loss function $L(x, y) =$	Domain
Bregman loss	$\phi'(\eta_2) - \phi'(\eta_1)$	$\phi(y) - \phi(x) - \phi'(x)(y - x)$	I
Squared error	$\eta_2 - \eta_1$	$(x - y)^2$	$\mathbb{R}$
Exponential Bregman	$\exp(\eta_2) - \exp(\eta_1)$	$\exp(y) - \exp(x) - \exp(x)(y - x)$	$\mathbb{R}$
QLIKE loss	$-1/\eta_2 + 1/\eta_1$	$y/x - \log(y/x) - 1$	$(0, \infty)$

Examples of members of the class $\mathcal{S}$ for the expectation functional, i.e., $V(x, y) = x - y$, are given in Table 1.2. While these examples are differentiable convex losses and therefore already covered in the literature (Luss and Rosset, 2014), the analysis in this paper also holds for the absolute loss, a nondifferentiable convex loss that is recovered when choosing $V(x, y) = \mathbb{1}\{x > y\} - 1/2$ and $\mathrm{d}H(\eta) = \mathrm{d}2\eta$. And even the elementary loss functions themselves bear relevance to fundamental decision problems in practice (Ehm et al., 2016). For the expectation functional, the elementary losses are nondifferentiable and convex, but describe the scenario of investing a fixed sum η for an unknown future profit or loss. For quantiles, the losses are not even convex, but describe the scenario of a bet on whether or not the outcome y will exceed the threshold η, with a fixed payoff ratio. While loss functions with properties such as convexity or differentiability are often

necessary in optimization problems for estimation, consumers of predictions regularly face decision problems with simpler loss structures. The results in this paper show that a distinction of preferences for technical implementation and forecast consumption is unnecessary in nonparametric isotonic regression.

1.3 Simultaneous optimality

Consider a distribution P for a random vector $(Z, Y) \in \mathcal{Z} \times \mathbb{R}$. We aim to minimize the criterion

$$\mathbb{E}_P S_\eta(g(Z), Y) \quad \text{for all } \eta \in \mathbb{R}, \tag{1.5}$$

over a family of regression functions $g\colon \mathcal{Z} \to \mathbb{R}$, and call a solution $\hat{g}$ simultaneously optimal since it minimizes the expected score with respect to all scoring functions in the class $\mathcal{S}$ at (1.4), simultaneously. Condition (1.5) is equivalent to minimizing $\mathbb{E}_P \mathbb{1}\{\eta \leq g(Z)\} V(\eta, Y)$ for all $\eta \in \mathbb{R}$. The results in this paper rely on this reformulation and the implication that regression functions are characterized by superlevel sets of the form $\{z \in \mathcal{Z}\colon \hat{g}(z) \geq \eta\}$, $\eta \in \mathbb{R}$. The structure of the set of admissible superlevel sets is crucial for the existence of a simultaneously optimal regression function.

In fact, it is a rare property in regression methods, that the solution does not depend on the loss function when considering a large class such as $\mathcal{S}$. As recently demonstrated, the optimal parameters with respect to the Bregman-loss criterion (1.3) of a parametric model $\{g_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$ of increasing functions g_θ generally vary depending on the chosen loss function (Patton, 2020). Before proving the simultaneous-optimality result for nonparametric isotonic regression in Section 1.4, we highlight the fragility of simultaneous optimality by demonstrating that it fails to hold for only slightly adapted shape constraints.

Unimodality is a shape constraint closely related to isotonicity. Given a predetermined mode, unimodality is even equivalent to isotonicity, when order relationships are defined suitably. For example, a total order on a finite set becomes a partial order consisting of two separate total orders merging in the predetermined mode, when reframing unimodality as isotonicity. Then, the problem becomes one of reconciling two isotonicity constraints. However, we will now see that simultaneous optimality under the unimodality constraint is in general unattainable when the location of the mode is not predetermined.

Example 1.3.1. Suppose that we have observations $(z_1, y_1), \ldots, (z_4, y_4)$ with $z_1 < \cdots < z_4$ and $(y_1, \ldots, y_4) = (9, 9, 0, 10)$, and let P denote the corresponding empirical distribution. We choose the expectation functional as the regression target, and for each potential mode $m_i = z_i$, $i = 1, \ldots, 4$, we aim to find a function $\hat{g}_i\colon \{z_1, \ldots, z_4\} \to \mathbb{R}$ that is optimal for any consistent loss function for the expectation functional. To this end, we reframe unimodality given a predetermined mode as isotonicity. The existence and the

uniqueness of an optimal isotonic solution for a functional of singleton type is shown in Section 1.4.

Using the PAV algorithm, the functions $\hat{g}_1$ and $\hat{g}_4$ are easy to find, as the order on the z_i is reversed or remains unchanged, respectively, when reframing the unimodality constraint as isotonicity. We refer to Section 1.4.2 and extant literature for a description of the algorithm. To find $\hat{g}_3$, we consider the partial order given by the totally ordered subsets $z_1 < z_2 < z_3$ and $z_3 > z_4$, and argue with superlevel sets of the form $\{z\colon \hat{g}_3(z) \geq \eta\}$, $\eta \in \mathbb{R}$. Since z_4 corresponds to the largest response in the data set, y_4, and z_3 needs to be in every nonempty superlevel set, we have $\hat{g}_3(z_3) = \hat{g}_3(z_4)$. Therefore, z_4 also lies in any nonempty superlevel set of $\hat{g}_3$, and in satisfying the isotonic relationship on $z_1 < z_2 < z_3$, we find that the only nonempty superlevel set must be $\{z_1, \ldots, z_4\}$, corresponding to levels $\eta \leq \frac{1}{4}\sum_{i=1}^4 y_i = 7$. Similarly, in order to find $\hat{g}_2$ as the isotonic solution subject to $z_1 < z_2$ and $z_2 > z_3 > z_4$, we again have $\hat{g}_2(z_3) = \hat{g}_2(z_4)$ since y_4 is the largest response. As $\frac{1}{2}\sum_{i=3}^4 y_i < y_2 = y_1$, isotonicity is established, and the only nonempty superlevel sets are $\{z_1, z_2\}$ and $\{z_1, \ldots, z_4\}$, corresponding to levels $\eta \in (5, 9]$ and $\eta \leq 5$, respectively. Coincidentally, $\hat{g}_2 = \hat{g}_1$.

The left panel of Figure 1.2 shows the regression functions and the right panel shows the expected score at (1.5) as a function of $\eta \in \mathbb{R}$. None of the three potential solutions minimizes the expected score for all η, and therefore a simultaneously optimal solution does not exist in this example. This visual method of comparing forecasts is called a Murphy diagram (Ehm et al., 2016).

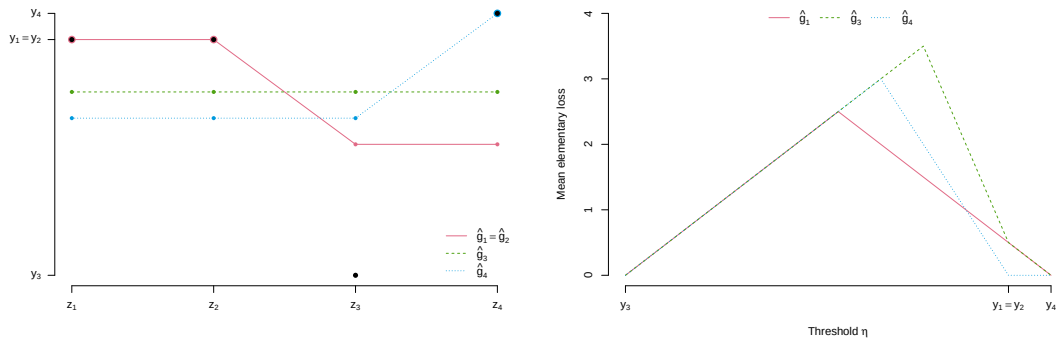

Figure 1.2: **Unimodal Regression and Murphy Diagram.** For a data example with observations $(z_1, 9), (z_2, 9), (z_3, 0), (z_4, 10)$, the left panel shows the regression functions $\hat{g}_1, \ldots, \hat{g}_4$ corresponding to modes $z_1, \ldots, z_4$. The black dots display the observations. The right panel shows the mean elementary losses of the regression functions against the parameter $\eta \in \mathbb{R}$. No single function exhibits the smallest mean elementary loss for all values of η, simultaneously.

In unimodal regression, a simultaneously optimal solution may but need not exist. This agrees with our findings in Section 1.4 because the set of admissible superlevel sets under

a unimodality shape constraint is not closed under union and intersection. Indeed, in Example 1.3.1 the sets $\{z_1\}$ and $\{z_4\}$ are admissible superlevel sets, while the union $\{z_1, z_4\}$ is not admissible because it implies bimodality.

1.4 Results on isotonic regression

We solve the isotonic regression problem considering a distribution P for a random vector $(Z, Y) \in \mathcal{Z} \times \mathbb{R}$, where $\mathcal{Z}$ is a finite partially ordered set. Analogously to (1.5), we aim to minimize the criterion

$$\mathbb{E}_P S_\eta(g(Z), Y) \quad \text{for all } \eta \in \mathbb{R}, \tag{1.6}$$

over all increasing functions $g\colon \mathcal{Z} \to \bar{\mathbb{R}}$. We call any minimizer of (1.6) a solution to the isotonic regression problem.

Reformulation of condition (1.6) as minimizing $\mathbb{E}_P \mathbb{1}\{\eta \le g(Z)\} V(\eta, Y)$ for all $\eta \in \mathbb{R}$ reveals that we can specify a solution to the isotonic regression problem by finding a path through minimizing upper sets $\{z \in \mathcal{Z}\colon \hat{g}(z) \ge \eta\}$. These upper sets are denoted by $x \in \mathcal{X} \subseteq \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}$ denotes the power set. The set $\mathcal{X}$ consists of all admissible superlevel sets for an increasing function g imposed by the partial order on $\mathcal{Z}$. A set $x \in \mathcal{X}$ is characterized by the property that if $z \in x$ and $z \preceq z'$, then $z' \in x$. This implies that $\mathcal{X}$ is a finite lattice, that is, it is closed under union and intersection and contains $\mathcal{Z}$ and the empty set. We will see, that as η increases, ξ follows one of the totally ordered paths through the lattice. In Figure 1.3 the direction of movement as η increases is illustrated by arrows. In the special case of a total order, $z_1 < \cdots < z_n$, there is only one possible path along upper sets of the form $\{z_i, \dots, z_n\}$, $i = 1, \dots, n$, ending up at the empty set.

The path is given by a function $\xi\colon \mathbb{R} \to \mathcal{X}$, that maps η to an upper set x of $\mathcal{Z}$ that minimizes

$$s_x(\eta) = v_x(\eta) = V(\eta, P_x) = \int_{x \times \mathbb{R}} V(\eta, y)\, P(\mathrm{d}z, \mathrm{d}y), \tag{1.7}$$

where $P_x(A) = P((x \times \mathbb{R}) \cap A)$ for any $A \in \mathcal{P}(\mathcal{Z}) \otimes \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. In this notation, s_x is only defined for $x \in \mathcal{X}$, whereas v_x and P_x are defined for any $x \in \mathcal{P}(\mathcal{Z})$. Again, we assume that all relevant integrals exist. For the bounds of the conditional functional, we write $T_x^- = T_{P_x}^- = \inf T(P_x)$ and $T_x^+ = T_{P_x}^+ = \sup T(P_x)$. Finally, let $X(\eta)$ denote the set of superlevel sets $x \in \mathcal{X}$ minimizing $s_x(\eta)$ at (1.7). Since $\mathcal{P}(\mathcal{Z})$ is finite, such a minimizer always exists.

For a total order, upper sets $\{z_i, \dots, z_n\}$ can be parameterized by the index of the smallest element, with the index $n+1$ for the empty set. Then we can redefine the object

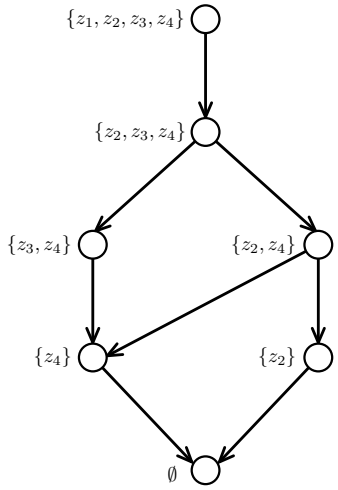

Figure 1.3: **Moving through the lattice $\mathcal{X}$.** The display shows possible paths through $\mathcal{X}$ based on the partial order on $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ given by $z_1 \prec z_2$ and $z_1 \prec z_3 \prec z_4$. The arrows indicate the direction of moving through the lattice $\mathcal{X}$ as η increases.

of minimization in (1.7) as

$$s_i(\eta) = \sum_{\ell=i}^{n} V(\eta, y_\ell).$$

This index search needs to be conducted for every $\eta \in \mathbb{R}$ separately. In Figure 1.4 we give an example for 6 data points. The example illustrates how the values $\hat{g}(z_\ell)$, $\ell = 1, \ldots, 6$, can be determined from the epigraph of the function $\eta \mapsto \min \xi(\eta)$. In a nutshell, for a total order, we find the generalized inverse to an optimal solution.

The following proposition formalizes that statement in the general context, assuming the existence of a decreasing function $\xi \colon \mathbb{R} \to \mathcal{X}$ in the sense that for $\eta' > \eta$ it holds that $\xi(\eta') \subseteq \xi(\eta)$, while satisfying $\xi(\eta) \in X(\eta)$ for all $\eta \in \mathbb{R}$. Before showing the existence of such a function ξ in Lemma 1.4.5, we elucidate the one-to-one correspondence to the solutions $\hat{g}$ of the isotonic regression problem at (1.6).

Proposition 1.4.1. *Let $\xi \colon \mathbb{R} \to \mathcal{X}$ be a decreasing, left-continuous function such that $\xi(\eta) \in X(\eta)$, where left-continuity means that if $\eta_n \uparrow \eta$ and $z \in \xi(\eta_n)$, then $z \in \xi(\eta)$. Then, the function $\hat{g} \colon \mathcal{Z} \to \mathbb{R}$ given by*

$$\inf\{\eta : z \notin \xi(\eta)\} = \hat{g}(z) = \max\{\eta : z \in \xi(\eta)\} \tag{1.8}$$

is the unique function that satisfies

$$\{z : g(z) \geq \eta\} = \xi(\eta) \quad \text{for all } \eta \in \mathbb{R},$$

among all increasing functions $g \colon \mathcal{Z} \to \mathbb{R}$.

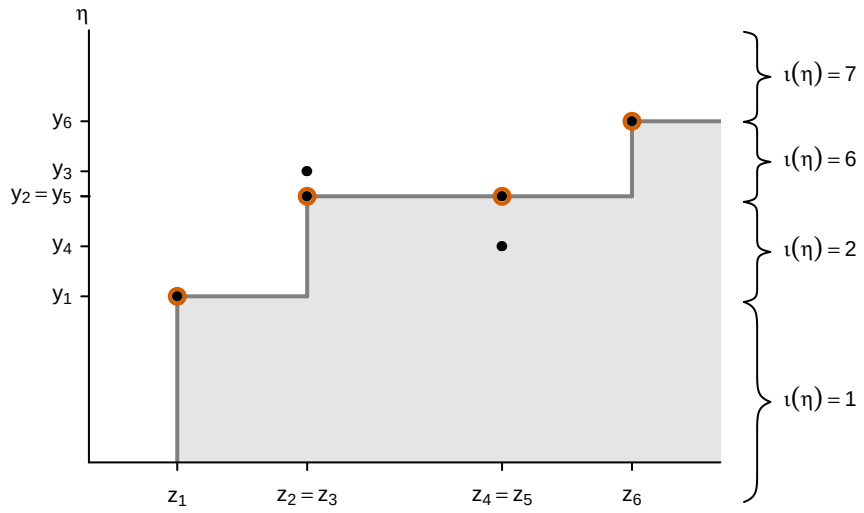

Figure 1.4: **Graph of $\hat{g}$.** For a sample of 6 data points with a totally ordered covariate set $\mathcal{Z}$, the values of $\hat{g}(z)$ for $z = z_1, \ldots, z_6$ are shown in red. The epigraph of the function $\eta \mapsto \min \xi(\eta) = z_{\iota(\eta)}$ is shown in grey, where T is chosen as the median functional to find $\xi(\eta)$.

Proof. The left-continuity and monotonicity of $\xi \colon \mathbb{R} \to \mathcal{X}$ implies the equality of infimum and maximum in equation (1.8). The monotonicity of $\hat{g}$ follows from the monotonicity of ξ and the fact that ξ takes values being superlevel sets of the partial order on $\mathcal{Z}$. Let $\eta' \in \mathbb{R}$. Then,

$$\text{(i)} \quad \hat{g}(z) \geq \eta' \implies \xi(\hat{g}(z)) \subseteq \xi(\eta') \implies z \in \xi(\eta').$$
$$\text{(ii)} \quad \text{For any } z \in \xi(\eta') : \quad \hat{g}(z) = \max\{\eta : z \in \xi(\eta)\} \geq \eta'.$$

Therefore, $\{z : \hat{g}(z) \geq \eta'\} \subseteq \{z : z \in \xi(\eta')\} \subseteq \{z : \hat{g}(z) \geq \eta'\}$ where the first inclusion follows by (i) and the second by (ii). Uniqueness follows because any hypothetical alternative $\bar{g}$ with $\bar{g}(z') \neq \hat{g}(z')$ for some $z' \in \mathcal{Z}$ leads to the contradiction $\xi(\eta) = \{z : \bar{g}(z) \geq \eta\} \neq \{z : \hat{g}(z) \geq \eta\} = \xi(\eta)$ for all η between $\bar{g}(z')$ and $\hat{g}(z')$. □

As a first result, we characterize minimizers of $s_x(\eta)$ at (1.7) for a given $\eta \in \mathbb{R}$. The following proposition states necessary and sufficient conditions for the inclusion of an upper set x in the set of minimizing superlevel sets $X(\eta)$. This is the first step towards establishing a link between the level η and the value of the functional T on the corresponding level set, and more elementary, it is also the first step in proving the existence of a decreasing function ξ as specified in Proposition 1.4.1.

Proposition 1.4.2. *Let $\eta \in \mathbb{R}$. Subject to $x, x' \in \mathcal{X}$, the inclusion $x \in X(\eta)$ holds if and only if*

$$v_{x\setminus x'}(\eta) \leq 0 \quad \text{for all } x' \subsetneq x,$$
$$v_{x'\setminus x}(\eta) \geq 0 \quad \text{for all } x' \supsetneq x.$$

Let $x \in X(\eta)$, $x' \in \mathcal{X}$. If $v_{x\setminus x'}(\eta) = v_{x'\setminus x}(\eta)$, then $x' \in X(\eta)$.

Proof. Note that $s_x(\eta) \leq s_{x'}(\eta)$ for all $x' \subsetneq x$ and all $x' \supsetneq x$ holds if and only if $v_{x\setminus x'}(\eta) \leq 0$ for all $x' \subsetneq x$ and $v_{x'\setminus x}(\eta) \geq 0$ for all $x' \supsetneq x$. For the first part of the result, note that $x \in X(\eta)$ implies $s_x(\eta) \leq s_{x'}(\eta)$ for all $x' \subsetneq x$ and all $x' \supsetneq x$. Conversely, let $x \in \mathcal{X}$ be such that the latter condition is satisfied. Then, $s_x(\eta) \leq s_{x'\cap x}(\eta)$ and $s_x(\eta) \leq s_{x'\cup x}(\eta)$ for all $x' \in \mathcal{X}$. By substracting $v_{x\setminus x'}(\eta)$ on both sides of the latter inequality, we have $s_{x\cap x'}(\eta) \leq s_{x'}(\eta)$ for all $x' \in \mathcal{X}$, and hence $x \in X(\eta)$. The second part of the result is immediate after adding $s_{x\cap x'}(\eta)$ to both sides of $v_{x\setminus x'}(\eta) = v_{x'\setminus x}(\eta)$, that is, $s_x(\eta) = s_{x'}(\eta)$. □

The following corollary is of particular importance in the context of total orders, where all admissible superlevel sets are pairwise nested.

Corollary 1.4.3. *Let $\eta \in \mathbb{R}$ and $x \in X(\eta)$, $x' \in \mathcal{X}$. If $x' \subsetneq x$ and $v_{x\setminus x'}(\eta) = 0$, then $x' \in X(\eta)$. Analogously, if $x' \supsetneq x$ and $v_{x'\setminus x}(\eta) = 0$, then $x' \in X(\eta)$.*

The next result establishes links between two or more sets of minimizing superlevel sets, that is, between $X(\eta)$ and $X(\eta')$ when $\eta \neq \eta'$. Afterwards, Lemma 1.4.5 shows the existence of a decreasing function ξ as specified in Proposition 1.4.1.

Lemma 1.4.4. *(a) Let $\eta, \eta' \in \mathbb{R}$, $\eta < \eta'$, and $x \in X(\eta)$, $x' \in X(\eta')$. Then, $v_{x'\setminus x}(\eta'') = 0$ for all $\eta'' \in [\eta, \eta']$.*

(b) Let $\eta \in \mathbb{R}$ and $x', x'' \in X(\eta)$, $x \in \mathcal{X}$. If $x \in \bigcup_{\eta\in\mathbb{R}} X(\eta)$ and $x' \supseteq x \supseteq x''$, then $x \in X(\eta)$.

(c) Let $\eta, \eta' \in \mathbb{R}$, $\eta < \eta'$, and $x \in X(\eta)$, $x' \in X(\eta')$. Then, $x \cup x' \in X(\eta)$ and $x \cap x' \in X(\eta')$.

Proof. (a) We have $(x \cup x') \setminus x = x' \setminus x = x' \setminus (x \cap x')$. The statement is trivial if $x' \setminus x = \emptyset$. Otherwise, $v_{x'\setminus x}(\eta) \geq 0 \geq v_{x'\setminus x}(\eta')$ by Proposition 1.4.2, where the statement follows from the monotonicity of the identification function in its first argument.

(b) The statement is trivial if $x = x'$, $x = x''$, or $x \notin X(\eta')$ for all $\eta' \neq \eta$. Therefore, assume $x \in X(\eta')$, $\eta' \neq \eta$. If $\eta < \eta'$, then $v_{x\setminus x''}(\eta) = 0$ by part (a). If $\eta' < \eta$, then $v_{x'\setminus x}(\eta) = 0$ by part (a). In either case, $x \in X(\eta)$ by Corollary 1.4.3.

(c) We have $s_x(\eta) \leq s_{x\cup x'}(\eta)$ and $s_{x'}(\eta') \leq s_{x\cap x'}(\eta')$, and $v_{x'\setminus x}(\eta'') = 0$ for all $\eta'' \in [\eta, \eta']$ by part (a). That means, $s_x(\eta) = s_{x\cup x'}(\eta)$ and $s_{x'}(\eta') = s_{x\cap x'}(\eta')$. □

Lemma 1.4.5. *(a) There exists a decreasing function $\xi\colon \mathbb{Q} \to \mathcal{X}$ such that $\xi(q) \in X(q)$ for all $q \in \mathbb{Q}$.*

(b) Let $\eta_n \uparrow \eta$ and $x_n \in X(\eta_n)$, $x_n \supseteq x_{n+1}$. Then, $x = \bigcap_{n\in\mathbb{N}} x_n \in X(\eta)$.

Proof. (a) Let $\{q_n\} = \mathbb{Q}$ be an enumeration of the rationals. We define $\xi(q_n)$ inductively. Pick $x_1 \in X(q_1)$ and set $\xi(q_1) = x_1$. For $n \geq 2$, define

$$x_n^- = \bigcup_{\substack{i\in\{1,\dots,n-1\}\\ q_i > q_n}} \xi(q_i), \quad x_n^+ = \bigcap_{\substack{i\in\{1,\dots,n-1\}\\ q_i < q_n}} \xi(q_i),$$

if $\{i : q_i > q_n\} \neq \emptyset$ and $\{i : q_i < q_n\} \neq \emptyset$. If $\{i : q_i > q_n\} = \emptyset$, we set $x_n^- = \emptyset$, and if $\{i : q_i < q_n\} = \emptyset$, we set $x_n^+ = \mathcal{Z}$. We choose any $x_n \in X(q_n)$ and set $\xi(q_n) = (x_n \cup x_n^-) \cap x_n^+$. At each step n, $\xi(q_n) \in X(q_n)$ follows by 1.4.4 (a), and $\xi(q_n) \subseteq x_n^+$. Furthermore, we show by induction that $x_n^- \subseteq \xi(q_n)$ for all n. For $n = 2$, this is easily verified. Suppose the claim holds for $n - 1 \geq 2$. If $q_n > q_{n-1}$, then $x_n^- = x_{n-1}^-$ and $x_n^+ = x_{n-1}^+ \cap \xi(q_{n-1}) = \xi(q_{n-1})$, hence

$$x_n^- = x_{n-1}^- \subseteq (x_n \cup x_{n-1}^-) \cap \xi(q_{n-1}) = \xi(q_n).$$

If $q_n < q_{n-1}$, then $x_n^- = x_{n-1}^- \cup \xi(q_{n-1}) = \xi(q_{n-1})$ and $x_n^+ = x_{n-1}^+$, hence

$$x_n^- = \xi(q_{n-1}) \subseteq (x_n \cup \xi(q_{n-1})) \cap x_{n-1}^+ = \xi(q_n).$$

In summary, for $k < n$, if $q_k < q_n$, then $\xi(q_n) \subseteq x_n^+ \subseteq \xi(q_k)$, and if $q_k > q_n$, $\xi(q_k) \subseteq x_n^- \subseteq \xi(q_n)$ showing that ξ is decreasing.

(b) We have $s_{x_n}(\eta_n) \leq s_{x'}(\eta_n)$ for all $x' \in \mathcal{X}$. Furthermore, the definitions of x and V imply $\mathbb{1}\{z \in x_n\}V(\eta_n, y) \to \mathbb{1}\{z \in x\}V(\eta, y)$ pointwise, and we have $\mathbb{1}\{z \in x_n\}V(\eta_n, y) \leq \sup_{n\in\mathbb{N}} |V(\eta_n, y)|$. By the dominated convergence theorem, $s_{x_n}(\eta_n) \to s_x(\eta)$ and $s_{x'}(\eta_n) \to s_{x'}(\eta)$. □

Part (b) of Lemma 1.4.5 describes a possible completion step for part (a) that also modifies ξ to be left-continuous. In a nutshell, any decreasing $\xi'\colon \mathbb{Q} \to \mathcal{X}$ that satisfies $\xi'(\eta') \in X(\eta')$ for all $\eta' \in \mathbb{Q}$ admits a left-continuous version on $\mathbb{R}$, $\xi : \eta \mapsto \bigcap_{\eta'<\eta} \xi'(\eta') \in X(\eta)$, where the intersection is over all $\eta' \in \mathbb{Q}$, $\eta' < \eta$.

In order to prove the existence of a function ξ (and thus $\hat{g}$) that solves the isotonic regression problem, we need that $\mathcal{X}$ is closed under union and intersection. This property is essential for Lemma 1.4.5.

We could also start with a set $\mathcal{X}$ of subsets of $\{z_1, \ldots, z_n\}$ that are interpreted as the admissible superlevel sets of the function g that is to be fitted. If $\mathcal{X}$ is closed under union and intersection, then $\mathcal{X}$ induces a partial order on $\{z_1, \ldots, z_n\}$ by Birkhoff's Representation Theorem; see for example Gurney and Griffin (2011). Consequently, the optimal function $\hat{g}$ always exists and is increasing.

Starting with $\mathcal{X}$, one could formulate constraints other than isotonicity on g as long as they can be formulated in terms of restrictions on admissible superlevel sets. Examples are unimodality or quasi-convexity. Generally, there is no solution that is simultaneously optimal with respect to all elementary loss functions; see Section 1.3 for an example in the case of a unimodality constraint.

1.4.1 Characterization of optimal solutions

The following proposition is essential to provide min-max and max-min bounds on solutions to the isotonic regression problem. We relate the threshold $\eta \in \mathbb{R}$ to the bounds of the functional T on subsets of the data. As a reminder, we write $T_x^- = T_{P_x}^- = \inf T(P_x)$ and $T_x^+ = T_{P_x}^+ = \sup T(P_x)$.

Proposition 1.4.6. *Let $\eta \in \mathbb{R}$, $x \in X(\eta)$. Then, subject to $x' \in \mathcal{X}$,*

$$\max_{x' \supsetneq x} T^-_{x' \setminus x} \leq \eta \leq \min_{x' \subsetneq x} T^+_{x \setminus x'},$$
$$\max_{x' \supsetneq x, x' \notin X(\eta)} T^+_{x' \setminus x} < \eta \leq \min_{x' \subsetneq x, x' \notin X(\eta)} T^-_{x \setminus x'}.$$

Proof. For all $x' \supsetneq x$, we have $v_{x' \setminus x}(\eta) \geq 0$. For all $x' \subsetneq x$, we have $v_{x \setminus x'}(\eta) \leq 0$. If $x' \notin X(\eta)$, then both inequalities are strict. Corollary 1.2.4 implies the result. □

Figure 1.5 illustrates the statement in Proposition 1.4.6 for a total order in the context of the expectation functional, which is a functional of singleton type. We now state and show one of our main results which is that $\hat{g}$ coincides with or is bounded by a min-max and max-min solution.

Theorem 1.4.7. *Let $z \in \mathcal{Z}$ and let $\hat{g}$ be a solution to the isotonic regression problem. Then, subject to $x, x' \in \mathcal{X}$,*

$$\min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-_{x \setminus x'} \leq \hat{g}(z) \leq \max_{x : z \in x} \min_{x' \subsetneq x} T^+_{x \setminus x'}.$$

Proof. Applying the first set of bounds from Proposition 1.4.6 to the formula for $\hat{g}$ at (1.8), we obtain

$$\inf_{\eta : z \notin \xi(\eta)} \max_{x \supsetneq \xi(\eta)} T^-_{x \setminus \xi(\eta)} \leq \hat{g}(z) \leq \max_{\eta : z \in \xi(\eta)} \min_{x' \subsetneq \xi(\eta)} T^+_{\xi(\eta) \setminus x'}.$$

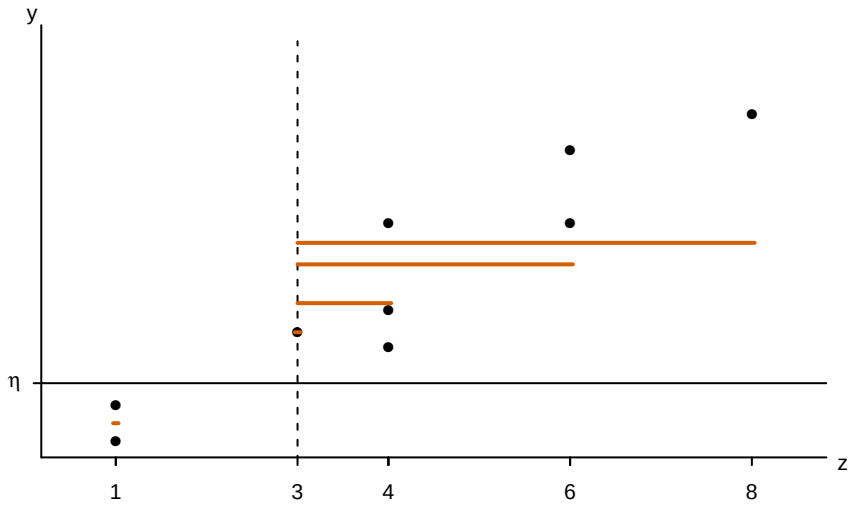

Figure 1.5: **Minimizing indices are separators.** For a sample of 9 data points, the graph illustrates the functional value (expectation) on relevant subsets of the data for a given η and the minimizing index $i = 3$. The expectation value (vertical location of a brown line) is above or below η when the corresponding subsample extends (horizontal extension of a brown line) to the right or left of the minimizing index, respectively.

The lower bound is bounded from below by $\min_{x':z\notin x'} \max_{x\supsetneq x'} T^-_{x\setminus x'}$, and the upper bound is bounded from above by $\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'}$. □

The previous statement is closely related to the coinciding max-min and min-max solutions at (1.2) for the expectation functional and a total order isotonicity constraint. For an analogous statement of uniqueness, as referred to in Example 1.3.1, we need the following lemma on a modified max-min inequality in the context of partial orders.

Lemma 1.4.8. *Suppose that T is of singleton type. Let $z \in \mathcal{Z}$ be such that $P(\{z\}\times\mathbb{R}) > 0$. Then, subject to $x, x' \in \mathcal{X}$,*

$$\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'} \le \min_{x':z\notin x'} \max_{x\supsetneq x'} T^-_{x\setminus x'}.$$

Proof. Let $x'' \in \mathcal{X}$ such that $z \notin x''$, then

$$\begin{aligned}\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'} &= \max_{x:z\in x} \min_{\substack{x'\subsetneq x \\ P((x\setminus x')\times\mathbb{R})>0}} T^+_{x\setminus x'} = \max_{x:z\in x} \min_{\substack{x'\subsetneq x \\ P((x\setminus x')\times\mathbb{R})>0}} T^-_{x\setminus x'} \\ &\le \max_{x:z\in x} T^-_{x\setminus(x\cap x'')} = \max_{x:z\in x} T^-_{(x\cup x'')\setminus x''} \le \max_{x:x\supsetneq x''} T^-_{x\setminus x''},\end{aligned}$$

where the last inequality holds because $x \cup x'' \in \mathcal{X}$ and if $z \in x$ then $x \cup x'' \supsetneq x''$. □

In general, a similar statement always holds, where the choice of ξ determines whether $\hat{g}$ attains the minimal or maximal elements of the functional. It is possible to define minimal and maximal solutions. Recall that we defined $X(\eta)$ as the set of superlevel sets $x \in \mathcal{X}$ minimizing $s_x(\eta)$ at (1.7). Let

$$X^-(\eta) = \{x \in X(\eta) : \nexists\, x' \in X(\eta) \text{ such that } x' \subsetneq x\},$$
$$X^+(\eta) = \{x \in X(\eta) : \nexists\, x' \in X(\eta) \text{ such that } x' \supsetneq x\}$$

denote the sets of minimal and maximal elements of $X(\eta)$, respectively.

Proposition 1.4.9. *Let $z \in \mathcal{Z}$ be such that $P(\{z\} \times \mathbb{R}) > 0$, and let $\xi \colon \mathbb{R} \to \mathcal{X}$ be decreasing and left-continuous.*

(a) If $\xi(\eta) \in X^+(\eta)$ for all $\eta \in \mathbb{R}$, then, subject to $x, x' \in \mathcal{X}$,

$$\hat{g}(z) = \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^+_{x \setminus x'} = \max_{x : z \in x} \min_{x' \subsetneq x} T^+_{x \setminus x'}.$$

(b) If $\xi(\eta) \in X^-(\eta)$ for all $\eta \in \mathbb{R}$, then, subject to $x, x' \in \mathcal{X}$,

$$\hat{g}(z) = \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-_{x \setminus x'} = \max_{x : z \in x} \min_{x' \subsetneq x} T^-_{x \setminus x'}.$$

Proof. The proof follows using Lemma 1.4.8 and applying the same steps as in the proof of Theorem 1.4.7 to the second set of bounds in Proposition 1.4.6. □

Let us denote the solution in part (a) of Proposition 1.4.9 by g^+ and the one in part (b) by g^-. Clearly, it always holds that $g^- \leq g^+$. It is a natural question whether any increasing function g that satisfies $g^- \leq g \leq g^+$ is also a minimizer of the criterion (1.6). It turns out that the answer is negative; see Mösching and Dümbgen (2020, Remark 2.2, Example 2.4). Combining Propositions 1.4.9 to 1.4.14 and Corollary 1.4.11, gives a complete characterizations of all possible solutions to the isotonic regression problem for partial orders. For the following results, it is not required that g^-, g^+ are the solutions from Proposition 1.4.9. Unless specified, they do not even need to satisfy $g^- \leq g^+$ everywhere. We define $\xi^- \colon \eta \mapsto \{z : g^-(z) \geq \eta\}$ and ξ^+ analogously.

Proposition 1.4.10. *Let g^- and g^+ be two solutions to the isotonic regression problem such that $g^- \leq g^+$. Let $\hat{g}$ be isotonic, $g^- \leq \hat{g} \leq g^+$, and suppose that all superlevel sets of $\hat{g}$ lie in $\bigcup_{\eta \in \mathbb{R}} X(\eta)$. Then, $\hat{g}$ is a solution to the isotonic regression problem.*

Proof. For $\eta \in \mathbb{R}$ define $\xi(\eta) = \{z : \hat{g}(z) \geq \eta\}$. The functions ξ, ξ^-, ξ^+ are decreasing, that is $\xi(\eta) \supseteq \xi(\eta')$ for $\eta \leq \eta'$, and left-continuous. For ξ^-, ξ^+ it holds that $\xi^-(\eta)$,

$\xi^+(\eta) \in X(\eta)$. Since, for all $z \in \mathcal{Z}$, it holds that

$$\begin{aligned} g^-(z) = \max\{\eta : z \in \xi^-(\eta)\} \leq g(z) = \max\{\eta : z \in \xi(\eta)\} \\ \leq g^+(z) = \max\{\eta : z \in \xi^+(\eta)\}, \end{aligned}$$

we obtain $\xi^-(\eta) \subseteq \xi(\eta) \subseteq \xi^+(\eta)$ for all $\eta \in \mathbb{R}$. Lemma 1.4.4 (b) implies the result. □

The following corollary is an immediate consequence of Lemma 1.4.4 (c).

Corollary 1.4.11. *Let g^- and g^+ be two solutions to the isotonic regression problem. Then, the distributive lattice generated by ξ^- and ξ^+ is a subset of $\bigcup_{\eta\in\mathbb{R}} X(\eta)$.*

Having two solutions g^- and g^+ allows us to find all solutions to the isotonic regression problem with superlevel sets that lie in the lattice generated by ξ^- and ξ^+. Examples include solutions that transition from g^- to g^+ at a particular threshold η,

$$\hat{g}(z) = \begin{cases} g^+(z), & z \in \xi^+(\eta), \\ g^-(z), & \text{otherwise}, \end{cases}$$

or pointwise convex combinations of solutions with $\alpha \in (0, 1)$,

$$\hat{g}(z) = \alpha g^-(z) + (1 - \alpha) g^+(z).$$

In order to refine the lattice of minimizing upper sets from Corollary 1.4.11 with the purpose to characterize all solutions, we pose the question whether simple separation rules exist for the set difference of consecutive lattice elements. These sets necessarily take the form of the intersection of a level set of g^- and a level set of g^+, that is, sets of the form $\{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$. These rules do exist as we show in Propositions 1.4.13 and 1.4.14. First, we introduce the notion of a separation.

Definition 1.4.12. A *separation* of a set $Z \in \mathcal{P}(\mathcal{Z})$ is a collection of sets $Z_1, \ldots, Z_n \subseteq Z$ that are pairwise separated and satisfy $Z = \bigcup_{i=1}^n Z_i$. Two sets Z_i and Z_j are *separated* with respect to Z if for all $z' \in Z_i$ and $z'' \in Z_j$, there does not exist a finite sequence $(z_k)_{k=1,\ldots,m}$, $z_k \in Z$, $z_1 = z'$, $z_m = z''$ that for all $k = 1, \ldots, m-1$ satisfies $z_k \preceq z_{k+1}$ or $z_{k+1} \preceq z_k$.

Proposition 1.4.13. *Let g^- and g^+ be two solutions to the isotonic regression problem, and let $\eta^-, \eta^+ \in \mathbb{R}$, $\eta^- < \eta^+$, be such that $Z = \{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$ is nonempty. Furthermore, let $Z_1, \ldots, Z_n$ be a separation of Z, and let $x' = \xi^-(\eta^-) \cap \xi^+(\eta^+)$ and $x'' = x' \setminus Z$. Then, $x'' \cup Z_k \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$, $k = 1, \ldots, n$.*

Proof. Without loss of generality, we show the claim for $k = 1$. By Lemma 1.4.4 (c), we have $x' \in X(\eta^+)$ and $x'' = \xi^-(\eta^- + \epsilon_1) \cup \xi^+(\eta^+ + \epsilon_2) \in X(\eta^- + \epsilon_1)$ for some $\epsilon_1, \epsilon_2 > 0$.

More precisely, we have $x', x'' \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ by Lemma 1.4.4 (b), since $\xi^-(\eta) \subseteq x'' \subseteq x' \subseteq \xi^+(\eta)$, $\eta \in (\eta^-, \eta^+]$.

Let $x_1 = x'' \cup Z_1$ and $x_2 = x' \setminus Z_1$ both of which are upper sets in $\mathcal{X}$. Then $Z_1 = x_1 \setminus x''$ but also $Z_1 = x' \setminus x_2$. Therefore, $v_{Z_1}(\eta) \geq 0 \geq v_{Z_1}(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ by Proposition 1.4.2. Then the statement follows from Corollary 1.4.3. □

Proposition 1.4.13 allows us to find additional solutions to the isotonic regression problem with superlevel sets where separation elements have been added to known minimizing superlevel sets. Using the variables defined in Proposition 1.4.13, one example of a new solution is

$$\hat{g}(z) = \begin{cases} \eta, & z \in Z_1, \\ g^+(z), & z \in x'', \\ g^-(z), & \text{otherwise}, \end{cases}$$

where $\eta \in (\eta^-, \eta^+]$. Iterative application of Proposition 1.4.13 recovers all minimizing superlevel sets that can be obtained from the solutions in Proposition 1.4.9 via Corollary 1.4.11 and the information on the partially ordered set $\mathcal{Z}$.

Proposition 1.4.14 allows us to recover the remaining minimizing superlevel sets when the distribution P of the random vector (Z, Y) is fully known. In fact, this proposition is a generalization of Proposition 1.4.13 that determines whether a level set intersection of g^- and g^+ can be split further by calculating values of the lower bound of the functional T.

Proposition 1.4.14. *Let g^- and g^+ be two solutions to the isotonic regression problem, and let $\eta^-, \eta^+ \in \mathbb{R}$, $\eta^- < \eta^+$, be such that $Z = \{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$ is nonempty. Furthermore, let $x' = \xi^-(\eta^-) \cap \xi^+(\eta^+)$ and $x'' = x' \setminus Z$. For $x \in \mathcal{X}$, $x' \supsetneq x \supsetneq x''$, we have $T^-_{x'\setminus x} \leq \eta^-$ if and only if $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$.*

Proof. We have $x', x'' \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ as in the proof of Proposition 1.4.13. Then, $v_{x'\setminus k}(\eta^+) \leq 0$ for all $k \in \mathcal{X}$, $k \subsetneq x'$, by Proposition 1.4.2, and hence $T^+_{x'\setminus k} \geq \eta^+$ by Corollary 1.2.4. Analogously, $v_{k\setminus x''}(\eta) \geq 0$ for all $k \in \mathcal{X}$, $k \supsetneq x''$, $\eta \in (\eta^-, \eta^+]$, leading to $T^-_{k\setminus x''} \leq \eta^-$.

For the first part of the statement, let $x \in \mathcal{X}$, $x' \supsetneq x \supsetneq x''$, be such that $T^-_{x'\setminus x} \leq \eta^-$. We show that $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ using Proposition 1.4.2. We have $T^+_{x'\setminus k} \leq \max\{T^-_{x'\setminus x}, T^+_{x\setminus k}\}$ for all $k \subsetneq x$ by Lemma 1.2.5. Since $T^-_{x'\setminus x} \leq \eta^-$ by assumption and as just shown $T^+_{x'\setminus k} \geq \eta^+$, we obtain $T^+_{x\setminus k} \geq \eta^+$. By Corollary 1.2.4, $v_{x\setminus k}(\eta) \leq 0$ for all $k \subsetneq x$, $\eta \leq \eta^+$, that is, the first inequality in Proposition 1.4.2 holds for all $\eta \in (\eta^-, \eta^+]$. Similarly, $T^-_{k\setminus x''} \geq \min\{T^-_{k\setminus x}, T^+_{x\setminus x''}\}$ for all $k \supsetneq x$. Since $T^-_{k\setminus x''} \leq \eta^-$ and $T^+_{x\setminus x''} \geq \eta^+$, we obtain $T^-_{k\setminus x} \leq \eta^-$. Therefore, $v_{k\setminus x}(\eta) \geq 0$, for all $\eta > \eta^-$, $k \supsetneq x$, that is, the second inequality in Proposition 1.4.2 holds for all $\eta \in (\eta^-, \eta^+]$.

To prove the converse, note that $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ implies that $v_{k\setminus x}(\eta) \geq 0$ for all $\eta \in (\eta^-, \eta^+]$, $k \supsetneq x$. Hence, in particular, $v_{x'\setminus x}(\eta) \geq 0$ and $T^-_{x'\setminus x} \leq \eta$ for all $\eta \in (\eta^-, \eta^+]$, and, therefore $T^-_{x'\setminus x} \leq \eta^-$. □

1.4.2 Pool-adjacent-violators algorithm

This section discusses the PAV algorithm and shows the optimality of its solution using the methods introduced in this paper. The algorithm solves the isotonic regression problem for a total order, taking observations $(z_1, y_1), \ldots, (z_n, y_n)$, $z_1 < \cdots < z_n$. Its starting point is the finest partition $\mathcal{Q}_0 = \{\{z_1\}, \ldots, \{z_n\}\}$ of the covariate set, and a corresponding function $g_0 \colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ satisfying

$$g_0(z) \in T(P_{\{z\}}).$$

If possible, an increasing function has to be chosen. The algorithm iteratively considers pooling adjacent elements Q_1 and Q_2 in the current partition, where "adjacent" means that the largest element of Q_1, $Q_1^+ = \max Q_1$, is the predecessor of the smallest element of Q_2, $Q_2^- = \min Q_2$. Pooling adjacent partition elements is considered necessary when $T^-_{Q_1} > T^+_{Q_2}$ (strong adjacent violators), it is considered invalid when $T^+_{Q_1} < T^-_{Q_2}$, and optional otherwise (weak adjacent violators). The early stopping criterion is the existence of an increasing function $g_{\mathrm{PAV}} \colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ that is constant on each element of the current partition $\mathcal{Q}_{\mathrm{PAV}}$ and satisfies

$$g_{\mathrm{PAV}}(z) \in T(P_Q) \quad \text{for all } Q \in \mathcal{Q}_{\mathrm{PAV}} \text{ and } z \in Q, \tag{1.9}$$

that is, when no further pooling is necessary. The late stopping criterion is reached when no weak adjacent violators remain. The first and most apparent property we observe is that for all $z \in \{z_1, \ldots, z_n\}$, $Q_1, Q_2 \in \mathcal{Q}_{\mathrm{PAV}}$, $Q_1^- \leq z \leq Q_2^+$, we have

$$T^-_{Q_1} \leq g_{\mathrm{PAV}}(z) \leq T^+_{Q_2}, \tag{1.10}$$

since otherwise either g_{PAV} is not increasing or the condition (1.9) is violated. Definition 1.2.1 and its Corollary 1.2.4 allow for an immediate proof of an additional property of $\mathcal{Q}_{\mathrm{PAV}}$.

Proposition 1.4.15. *Let $\mathcal{Q}$ be a partition of $\{z_1, \ldots, z_n\}$ found by the PAV algorithm, $Q \in \mathcal{Q}$, and $z \in Q$. Then,*

$$T^-_{Q|_{\geq z}} \leq T^-_Q \leq T^+_Q \leq T^+_{Q|_{\leq z}},$$

where $Q|_{\geq z}$ and $Q|_{\leq z}$ denote the restrictions to the elements $q \in Q$ satisfying $q \geq z$ and $q \leq z$, respectively.

Proof. The second inequality is trivial. For the first inequality, suppose the contrary: There exist $\eta \in \mathbb{R}$, $z \in Q$ such that $T^-_Q < \eta < T^-_{Q|_{\geq z}}$. This implies that $Q \neq Q|_{\geq z}$ and $v_Q(\eta) \geq 0 > v_{Q|_{\geq z}}(\eta)$, hence $v_{Q|_{<z}}(\eta) > 0$. Therefore, $T^+_{Q|_{<z}} < \eta < T^-_{Q|_{\geq z}}$, which means that Q can be seen as the result of an invalid pooling of $Q|_{<z}$ and $Q|_{\geq z}$. A similar argument applies to the third inequality. □

To show the connection between a valid solution by the PAV algorithm and the score optimizing solution $\hat{g}$ in Section 1.4, we define

$$\xi_{\mathrm{PAV}}(\eta) = \{z : \eta \leq g_{\mathrm{PAV}}(z)\}, \tag{1.11}$$

which are necessarily sets of the form $\{z_i, \ldots, z_n\}$. Plugging ξ_{PAV} into the definition of $\hat{g}$ recovers g_{PAV},

$$\begin{aligned}\hat{g}(z) &= \max\{\eta : z \in \xi_{\mathrm{PAV}}(\eta)\} \\ &= \max\{\eta : \eta \leq g_{\mathrm{PAV}}(z)\} = g_{\mathrm{PAV}}(z).\end{aligned}$$

In order to show that g_{PAV} solves the isotonic regression problem, it remains to be shown that $\xi_{\mathrm{PAV}}(\eta) \in X(\eta)$ for all $\eta \in \mathbb{R}$.

Proposition 1.4.16. *Let $\eta \in \mathbb{R}$, then $\xi_{\mathrm{PAV}}(\eta) \in X(\eta)$.*

Proof. Let $\eta \in \mathbb{R}$ and $x = \xi_{\mathrm{PAV}}(\eta)$. For all $Q \in \mathcal{Q}_{\mathrm{PAV}}$, we have $T^-_{Q\setminus x} < \eta \leq T^+_{Q\cap x}$ by statement (1.10) and defining equality (1.11). Recall that $T^-_\emptyset$ and $T^+_\emptyset$ are $-\infty$ and ∞, respectively. We now use that $T^-_{P_1+P_2} \leq \max\{T^-_{P_1}, T^-_{P_2}\}$ and $T^+_{P_1+P_2} \geq \min\{T^+_{P_1}, T^+_{P_2}\}$ for nonnegative measures P_1 and P_2 on $\mathbb{R}$, which is an immediate consequence of Definition 1.2.1. Together with Proposition 1.4.15, and subject to x' denoting an upper set of the form $\{z_i, \ldots z_n\}$, we have $T^-_{x'\setminus x} \leq \max_{Q\in\mathcal{Q}_{\mathrm{PAV}}} T^-_{Q\setminus x}$ for all $x' \supsetneq x$, and $T^+_{x\setminus x'} \leq \min_{Q\in\mathcal{Q}_{\mathrm{PAV}}} T^+_{Q\cap x}$ for all $x' \subsetneq x$. Therefore, $v_{x'\setminus x}(\eta) \geq 0$ for all $x' \supsetneq x$, and $v_{x\setminus x'}(\eta) \leq 0$ for all $x' \subsetneq x$, and the statement follows from Proposition 1.4.2. □

As a closing side note, we point out that ξ_{PAV} corresponds to coarsest partition that allows the solution g_{PAV}. Any weak adjacent violators on which g_{PAV} takes the same value have been pooled.

1.4.3 Partitioning the covariate set

In Section 1.4.2, we discussed how the PAV algorithm creates a partition of $\mathcal{Z}$, and that it leads to a solution $\hat{g}$ of the isotonic regression problem in the context of total orders. In this section, we show how a solution to the isotonic regression problem leads to a

corresponding partition $\mathcal{Q}$ of $\mathcal{Z}$, such that the solution satisfies

$$\hat{g}(z) \in T(P_Q), \quad \text{for all } Q \in \mathcal{Q},\ z \in Q,$$

and the solution is constant on every element of the partition. Let T be a functional of singleton type, and $\hat{g}$ be a solution to the isotonic regression problem. Subject to $x, x', k, k' \in \mathcal{X}$, the combination of Theorem 1.4.7 and Lemma 1.4.8 yields

$$\begin{aligned}\hat{g}(z) &= \max_{x: z \in x} \min_{x' \subsetneq x} T^+_{x \setminus x'} \\ &= \min_{k': z \notin k'} \max_{k \supsetneq k'} T^-_{k \setminus k'}.\end{aligned}$$

for all $z \in \mathcal{Z}$ with $P(\{z\} \times \mathbb{R}) > 0$. We call (x, x') a max-min pair for z if $z \in x$, $x' \subsetneq x$, and $\hat{g}(z) = T^+_{x \setminus x'}$, and we call (k', k) a min-max pair for z if $z \notin k'$, $k \supsetneq k'$, and $\hat{g}(z) = T^-_{k \setminus k'}$. For a pair $x, x' \in \mathcal{X}$ such that $T^-_{x \setminus x'} = T^+_{x \setminus x'}$, we also use the notation $T^\pm_{x \setminus x'}$. Note that for a functional T of singleton type, we have $T(P_{x \setminus x'}) = \{T^\pm_{x \setminus x'}\}$ if $P((x \setminus x') \times \mathbb{R}) > 0$. The following lemma provides the necessary tools to construct the partition $\mathcal{Q}$.

Lemma 1.4.17. *Let T be a functional of singleton type, and $\hat{g}$ be a solution to the isotonic regression problem. Furthermore, let $z \in \mathcal{Z}$ such that $P(\{z\} \times \mathbb{R}) > 0$, and let $(x_1, x_1'), (x_2, x_2')$ be max-min pairs for z, and $(k_1', k_1), (k_2', k_2)$ be min-max pairs for z. Then the following statements hold:*

(a) We have that $\hat{g}(z) = T^\pm_{x_1 \setminus k_1'} = T^\pm_{(x_1 \cup x_2) \setminus k_1'} = T^\pm_{x_1 \setminus (k_1' \cap k_2')}$.

(b) If $x, k' \in \mathcal{X}$ such that $z \in x$, $z \notin k'$, and $\hat{g}(z) = T^\pm_{x \setminus k'}$, then $(x, x \cap k')$ is a max-min pair for z and $(k', k' \cup x)$ is a min-max pair for z.

(c) If $\tilde{z} \in x_1 \setminus k_1'$, then (x_1, x_1') is a max-min pair for $\tilde{z}$ and (k_1', k_1) is a min-max pair for $\tilde{z}$.

Proof. We repeatedly use the inequalities $\hat{g}(z) = T^+_{x_1 \setminus x_1'} = \min_{x' \in \mathcal{X}} T^+_{x_1 \setminus x'} \leq T^+_{x_1 \setminus k'}$ and $\hat{g}(z) = T^-_{k_1 \setminus k_1'} = \max_{k \in \mathcal{X}} T^-_{k \setminus k_1'} \geq T^-_{x \setminus k_1'}$ for all $x, k' \in \mathcal{X}$, where the second equality holds because $T^+_P = \infty$ and $T^-_P = -\infty$ for null measures P. Furthermore, by assumption, $T(P_{x \setminus k'})$ is a singleton if $P((x \setminus k') \times \mathbb{R}) > 0$, and therefore $T(P_{x \setminus k'})$ is a singleton if $z \in x$ and $z \notin k'$.

(a) Clearly, $z \in x_1$, $z \in x_2$, $z \notin k_1'$, and $z \notin k_2'$. Hence, $\hat{g}(z) \leq T^\pm_{x_1 \setminus k_1'} \leq \hat{g}(z)$ implies the first statement. Furthermore, $\hat{g}(z) \leq T^+_{x_2 \setminus (x_1 \cup k_1')} = T^+_{(x_2 \setminus x_1) \setminus k_1'}$, and hence $\hat{g}(z) = \min\{T^-_{x_1 \setminus k_1'}, T^+_{(x_2 \setminus x_1) \setminus k_1'}\} \leq T^\pm_{(x_1 \cup x_2) \setminus k_1'} \leq \hat{g}(z)$ confirms the second statement using Lemma 1.2.5. Similarly, for the third statement, $\hat{g}(z) \leq T^\pm_{x_1 \setminus (k_1' \cap k_2')} \leq \max\{T^+_{x_1 \setminus k_1'}, T^-_{(x_1 \cap k_1') \setminus k_2'}\} = \hat{g}(z)$.

(b) The statement follows immediately from $T^-_{x\setminus k'} = T^+_{x\setminus k'}$, $(x \cup k') \setminus k' = x \setminus k' = x \setminus (x \cap k')$, and the definition of max-min and min-max pairs.

(c) Let $(x_{\tilde{z}}, x'_{\tilde{z}})$ be a max-min pair for $\tilde{z}$ and $(k'_{\tilde{z}}, k_{\tilde{z}})$ be a min-max pair for $\tilde{z}$. Then the statement follows from $\hat{g}(z) \leq T^{\pm}_{x_1\setminus k'_{\tilde{z}}} \leq \hat{g}(\tilde{z}) \leq T^{\pm}_{x_{\tilde{z}}\setminus k'_1} \leq \hat{g}(z)$. □

Proposition 1.4.18. *Let T be a functional of singleton type. Then there exists a partition $\mathcal{Q}$ of $\mathcal{Z}$ such that $\hat{g}$ is constant on every element of the partition almost everywhere and $\hat{g}(z) \in T(P_Q)$ for all $Q \in \mathcal{Q}$, $z \in Q$ such that $P(\{z\} \times \mathbb{R}) > 0$.*

Proof. Let $\bar{x}_z$ denote the union of the first components of all max-min pairs for $z \in \mathcal{Z}$, and let $\bar{k}'_z$ denote the intersection of the first components of all min-max pairs for $z \in \mathcal{Z}$. By Lemma 1.4.17 (a), we have $\hat{g}(z) = T^{\pm}_{\bar{x}_z\setminus\bar{k}'_z}$. We now show that the collection $\mathcal{Q}$ of sets $Q_z = \bar{x}_z \setminus \bar{k}'_z$ is a partition of $\mathcal{Z}$. First, we have $\bigcup_{z\in\mathcal{Z}} Q_z = \mathcal{Z}$, since $z \in \bar{x}_z$ and $z \notin \bar{k}'_z$ for all $z \in \mathcal{Z}$. Second, by Lemma 1.4.17 (b), we have that $(\bar{x}_z, \bar{x}_z \cap \bar{k}'_z)$ is a max-min pair for z and $(\bar{k}'_z, \bar{k}'_z \cup \bar{x}_z)$ is a min-max pair for z. Then, by Lemma 1.4.17 (c), we have $\bar{x}_z \subset \bar{x}_{\tilde{z}}$ and $\bar{k}'_z \supset \bar{k}'_{\tilde{z}}$ for all $\tilde{z} \in Q_z$, i.e., $Q_z \subset Q_{\tilde{z}}$ and in particular $z \in Q_{\tilde{z}}$. Swapping the roles of z and $\tilde{z}$ gives $Q_{\tilde{z}} \subset Q_z$. Therefore, $Q_z = Q_{\tilde{z}}$ for all $z \in \mathcal{Z}, \tilde{z} \in Q_z$. □

When T is a functional of interval type, we therefore obtain a partition for every fixed convex combination of its lower bound T^- and its upper bound T^+.

2 Isotonic regression for functionals of elicitation complexity greater than one

Anja Mühlemann and Johanna F. Ziegel

Abstract. The solutions to the isotonic regression problem for one-dimensional elicitable functionals T have been thoroughly studied and are well-understood. Interestingly, these solutions are robust with respect to the choice of loss function, in the sense that no matter which strictly consistent loss function for the functional T is chosen, we will obtain the same isotonic solution. We call these solutions simultaneously optimal. However, not all functionals are elicitable. Prominent examples include the expected shortfall, an important risk measure in finance, and the variance. Although not elicitable themselves these and other non-elicitable functionals can be obtained as a function of a 2-dimensional elicitable functional. In this manuscript we solve the isotonic regression problem for bivariate functionals of this type. However, for bivariate functional the results differ from the univariate case. We show that simultaneous optimality with respect to an entire class of loss functions can rarely be achieved and show how the isotonic regression problem can be solved for specific choices of loss functions. We also introduce a natural criterion to check whether a solution is simultaneously optimal and examine our findings in a simulations study. Furthermore, we show how the results can be generalized to partially ordered covariate sets.

Acknowledgments. We would like to thank Alexander I. Jordan and Alexander Henzi for valuable discussions. Furthermore, we gratefully acknowledge financial support from the Swiss National Science Foundation.

2.1 Introduction

In isotonic regression the aim is to fit an increasing function g_1 to observations $(z_1, y_1), \ldots, (z_n, y_n)$ such that a chosen loss function is minimized by g_1. The solution g_1 is then called a solution to the isotonic regression problem. If g_1 is supposed to model a conditional mean, then the loss function should be *consistent* for the mean in the sense of Gneiting (2011, Definition 1) with a prominent example being the squared error loss. More generally, if g_1 is a model for a conditional functional T, then the loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ should be chosen *consistent* for this functional T, that is, $\mathbb{E}_P L(t, Y) \leq \mathbb{E}_P L(x, Y)$ for all relevant probability distributions P, all $t \in T(P)$ and all $x \in \mathbb{R}$. Loss L is *strictly consistent* if the above inequality is strict. The notion of consistency used in this paper is a property of the functional and the loss function and not to be confused with consistency of an estimator. Strict consistency of L ensures that a correctly specified model minimizes the expected loss at the population level.

If a functional T, that is, a map on a certain class of probability distributions, has a strictly consistent loss function it is called *elicitable*. We say that the loss function *elicits* T. Elicitability is important for forecast comparison (Gneiting, 2011), and yields natural estimation procedures. Unfortunately, some ubiquitous functionals are not elicitable with prominent examples being variance (var) and expected shortfall (ES_α), the latter being an important risk measure in finance and insurance. Interestingly, although ES_α is not elicitable, it is jointly elicitable together with the α-quantile (q_α); see Fissler and Ziegel (2016) and Example 2.2.2. Similarly, while var itself is not elicitable, it is jointly with the mean ($\mathbb{E}$). This means that both ES_α and var are 2-elicitable, that is, they can both be obtained as a function of a 2-dimensional elicitable functional. In a nutshell, the *elicitation complexity* of a functional is the minimal number k of dimensions needed for the functional to be k-elicitable. Since both ES_α and var are not elicitable themselves but 2-elicitable their elicitation complexity equals 2 (Frongillo and Kash, 2020, Corollary 1 and 3).

Isotonic regression for one-dimensional elicitable functionals is well-understood (Barlow et al., 1972). An interesting aspect is its robustness with respect to the choice of the consistent loss function in the minimization problem. In other words, no matter which strictly consistent loss function we choose for the functional T, we will obtain the same isotonic solution (Brümmer and Du Preez, 2013; Jordan et al., 2019). This is in stark contrast to estimation in parametric regression models. In finite samples or for misspecified models, the choice of the consistent loss function may lead to miscellaneous estimates (Patton, 2020) and Chapter 4.

In this article, we investigate non-parametric regression for bivariate functionals $\underline{T}$ under isotonicity constraints. In particular, we show that simultaneous optimality with respect to an entire class of losses can rarely be achieved, and discuss how to find optimal solutions for specific choices of loss functions.

The functionals we consider are of the form

$$\underline{T} = (T, \underline{L}),$$

where T is a one-dimensional elicitable functional with strictly consistent loss function L, and

$$\underline{L}(P) := \inf_{x_1 \in \mathbb{R}} L(x_1, P) \tag{2.1}$$

with $L(x_1, P) = \int_{-\infty}^{\infty} L(x_1, y)\, \mathrm{d}P(y)$ is the *Bayes risk*. The example $\underline{T} = (\mathbb{E}, \mathrm{var})$ arises by choosing $L(x, y) = (x - y)^2$, and the example $\underline{T} = (q_\alpha, \mathrm{ES}_\alpha)$ is obtained by choosing $L(x,y) = (1/\alpha)\mathbb{1}\{y \leq x\}(x-y)-x$, which is the piecewise linear loss known from quantile regression up to a function that only depends on y. Generally, Frongillo and Kash (2020) show that $\underline{T}$ is always 2-elicitable. Moreover, they also introduce a large class $\underline{\mathcal{L}}$ of loss functions $L(x_1, x_2, y)$ eliciting $\underline{T}$.

We show how the isotonic regression problem can be solved for $\underline{T}$. It turns out that the proposed canonical solution is generally not optimal with respect to all loss functions in $\underline{\mathcal{L}}$, but there is a fairly simple approach to check whether a given fit is simultaneously optimal. Furthermore, we show how the fit can be improved for a specific chosen loss function. In a simulation study how often simultaneously optimal fits occur for the functionals (q_α, ES) and $(\mathbb{E}, \mathrm{var})$ and investigate the fits for a specific choice of loss function.

The article is organized as follows. Section 2.2 introduces necessary preliminaries on consistent loss functions including a mixture representation for loss functions in $\underline{\mathcal{L}}$. In Section 2.3, the isotonic regression problem for total orders is formulated and a natural solution through sequential optimization is proposed. Then, we study the simultaneous optimality of the solution of the sequential optimization approach. Section 2.4 contains the simulation study. In the Appendix, we show how our results can be generalized to partial orders.

2.2 Preliminaries

Following Jordan et al. (2019), a function $V : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is called an *identification function* if $V(\cdot, y)$ is increasing and left-continuous for all $y \in \mathbb{R}$. Then, for any probability measure P on $\mathbb{R}$ with finite support, we define the *functional* T induced by an identification function V as

$$T(P) = [T^-(P), T^+(P)] \subseteq [-\infty, \infty],$$

where the lower and upper bounds are given by

$$T^-(P) = \sup\{x_1 : V(x_1, P) < 0\} \quad \text{and} \quad T^+(P) = \inf\{x_1 : V(x_1, P) > 0\},$$

using the notation $V(x_1, P) = \int_{-\infty}^{\infty} V(x_1, y)\,\mathrm{d}P(y)$. A broad class of functionals can be defined via their identification function, quantiles and expectiles, including the median and the mean, just being some of the most prominent examples. For other popular examples, see Jordan et al. (2019). The examples of quantiles and expectiles already illustrate that the functional T can take singleton-values as well as interval-values.

Theorem 1 in Frongillo and Kash (2020) states that if L is a strictly consistent loss function for T and $\underline{L}$ is the Bayes risk defined at (2.1), then the loss

$$\tilde{L}(x_1, x_2, y) = L'(x_1, y) + H(x_2) + h(x_2)(L(x_1, y) - x_2) \tag{2.2}$$

elicits $\underline{T} = (T, \underline{L})$, where $h : \mathbb{R} \to \mathbb{R}$ is any positive strictly decreasing function, $H(r) = \int_0^r h(x)\,\mathrm{d}x$, and L' is any other consistent loss function for T. If h is merely decreasing, then $\tilde{L}$ is still a consistent loss function.

Ehm et al. (2016) showed that for expectiles and quantiles any consistent loss function L' can be written as

$$L'(x_1, y) = \int_{\mathbb{R}} S_{\eta,1}(x_1, y)\,\mathrm{d}H_1(\eta), \tag{2.3}$$

for certain elementary (quantile or expectile) losses $S_{\eta,1}$ and a non-negative measure H_1 on $\mathbb{R}$ depending on L'. In fact, such mixtures always yield a large class $\mathcal{L}$ of consistent scoring functions for T if it is identifiable with identification function $V(x, y)$ (Dawid, 2016; Ziegel, 2016a). Then, the elementary losses are given by

$$S_{\eta,1}(x_1, y) = (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})\, V(\eta, y), \tag{2.4}$$

where $\eta \in \mathbb{R}$. Moreover, the elementary losses are themselves consistent for T. We define

$$\mathcal{L} = \left\{ (x_1, y) \mapsto \int_{\mathbb{R}} S_{\eta,1}(x_1, y)\,\mathrm{d}H_1(\eta) : H_1 \text{ is a positive measure on } \mathbb{R} \right\}.$$

Note that (strict) consistency of a loss function is not altered by adding functions in y as long as they are integrable for all relevant probability measures P. Therefore, when speaking of characterizations of class of (strictly) consistent loss functions this is always meant up to possible addition of functions in y.

If a loss function is given as a mixture of elementary losses as in (2.3), this may be useful when minimizing the expected loss (over some set of parameters, for example); see details for the isotonic regression problem in Section 2.3. Using Fubini's theorem, one can see that we can look for minimizers of the expected elementary losses and hope that these minimizers all agree, that is, there is a simultaneous minimizer for all parameters η. Then, this minimizer is automatically optimal for all scoring functions of the form (2.3), independently of the measure H_1. Indeed, this approach is at the heart of the characterization of all simultaneously optimal solutions to the isotonic regression problem for one-dimensional functionals in Jordan et al. (2019).

Using the same approach as used by Ziegel et al. (2020) to derive a mixture representation for the pair $(q_\alpha, \mathrm{ES}_\alpha)$, we derive a mixture representation for the loss functions for $\underline{T}$ of the form (2.2).

Lemma 2.2.1. *Let $L, L' \in \mathcal{L}$. Then, all consistent loss functions for $\underline{T} = (T, \underline{L})$ defined at (2.2) are of the form*

$$\tilde{L}(x_1, x_2, y) = \int S_{\eta,1}(x_1, y)\, \mathrm{d}H_1(\eta) + \int S_{\eta,2}(x_1, x_2, y)\, \mathrm{d}H_2(\eta), \tag{2.5}$$

where H_1, H_2 are non-negative measures on $\mathbb{R}$, H_2 is finite on intervals of the form $(-\infty, -x_2]$, $x_2 \in \mathbb{R}$, and

$$\begin{aligned} S_{\eta,1}(x_1, y) &= (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})V(\eta, y) \\ S_{\eta,2}(x_1, x_2, y) &= \mathbb{1}\{\eta \le -x_2\}(L(x_1, y) + \eta) - \mathbb{1}\{\eta \le 0\}\eta. \end{aligned}$$

Conversely, any loss function of the form (2.5) is consistent for $\underline{T} = (T, \underline{L})$. It is strictly consistent if H_2 puts positive mass on all open intervals.

Proof. The consistency follows directly from Theorem 1 in Frongillo and Kash (2020). Recall that h is decreasing and nonnegative and $H(r) = \int_0^r h(x)\, \mathrm{d}x$. To see that the loss functions in (2.2) with loss $L' \in \mathcal{L}$ can be written as in (2.5), define $A := \lim_{x\to\infty} h(x) \ge 0$. Since $h \ge 0$, we can define the measure H_2 by $H_2((-\infty, t]) = h(-t) - A \ge 0$ for all $t \in \mathbb{R}$. Without loss of generality we can assume h satisfies $\lim_{x\to\infty} h(x) = 0$. Indeed, we can define $\underline{h} = h - A$ then H becomes $\underline{H}(x) = H(x) - xA$ and $\tilde{L}(x_1, x_2, y) = \underline{\tilde{L}}(x_1, x_2, y) + AL(x_1, y)$. Then, $\underline{\tilde{L}}(x_1, x_2, y) = L'(x_1, y) - AL(x_1, y) + H(x_2) + h(x_2)(L(x_1, y) - x_2)$. Thus, adding constants to h corresponds to modifying the loss function L'. Moreover, since $L, L' \in \mathcal{L}$ we have that $L' + AL \in \mathcal{L}$. Hence, we can assume that $A = 0$, then $H_2((-\infty, x]) = h(-x)$ for all $x \in \mathbb{R}$ and

$$h(x_2) = \int_{-\infty}^{-x} \mathrm{d}H_2(\eta).$$

Then we have

$$\int S_{\eta,2}(x_1, x_2, y)\, \mathrm{d}H_2(\eta) = L(x_1, y)h(x_2) - \int_{-x_2}^{0} \eta\, \mathrm{d}H_2(\eta).$$

Integration by parts yields

$$\int S_{\eta,2}(x_1, x_2, y)\, \mathrm{d}H_2(\eta) = L(x_1, y)h(x_2) - x_2 h(x_2) + H(x_2).$$

Restricting the choice of L' to $\mathcal{L}$ ensures the existence of the mixture representation for $L'(x_1, y)$. □

The following two examples discuss the mixture representations for the pairs $(q_\alpha, \mathrm{ES}_\alpha)$ and $(\mathbb{E}, \mathrm{var})$ in more detail.

Example 2.2.2. As already mentioned in the introduction a popular but non-elicitable risk measure is the expected shortfall. While in Fissler and Ziegel (2016) and Ziegel et al. (2020) ES_α typically takes negative values, Frongillo and Kash (2020) opted for a different sign convention so that ES_α is typically positive. In Fissler and Ziegel (2016) and Ziegel et al. (2020), the values of ES_α represent the gain that could be made. Thus, negative values correspond to a negative gain, that is, a loss that could be incurred with a certain transaction. In Frongillo and Kash (2020), the values of ES_α stand for the absolute amount of loss that one could face, while negative values correspond to a negative loss, or in other words, a gain. In this article we adopt the sign convention used by Frongillo and Kash (2020).

For a given level $\alpha \in (0,1)$ the loss function

$$L(x_1, y) = \frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y) - x_1$$

elicits the α-quantile $q_\alpha(P)$. The expected shortfall ES_α is the corresponding Bayes risk, that is,

$$\mathrm{ES}_\alpha(P) = \inf_{x_1 \in \mathbb{R}} L(x_1, P).$$

The elementary loss functions of Lemma 2.2.1 are given by

$$\begin{aligned} S_{\eta,1}(x_1, y) &= (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})(\mathbb{1}\{\eta > y\} - \alpha) \\ S_{\eta,2}(x_1, x_2, y) &= \mathbb{1}\{\eta \le -x_2\}\left(\tfrac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y) - (x_1 - \eta)\right) - \mathbb{1}\{\eta \le 0\}\eta. \end{aligned}$$

In fact, all loss functions consistent for the pair $(q_\alpha, \mathrm{ES}_\alpha)$ are of the form (2.2), or equivalently, (2.5); see Ziegel et al. (2020). Due to the different sign conventions mentioned previously, the mixture representation in Ziegel et al. (2020) corresponds to $L(x_1, -x_2, y)$ (up to normalization).

Example 2.2.3. The squared loss $L(x_1, y) = (x_1 - y)^2$ elicits the expectation $\mathbb{E}(P)$. The corresponding Bayes risk is the variance $\mathrm{var}(P)$. Thus, the pair $(\mathbb{E}, \mathrm{var})$ is elicitable. The elementary loss functions of Lemma 2.2.1 are given by

$$\begin{aligned} S_{\eta,1}(x_1, y) &= (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})(\eta - y) \\ S_{\eta,2}(x_1, x_2, y) &= \mathbb{1}\{\eta \le -x_2\}\left((x_1 - y)^2 + \eta)\right) - \mathbb{1}\{\eta \le 0\}\eta. \end{aligned}$$

In contrast to the pair $(q_\alpha, \mathrm{ES}_\alpha)$ not all consistent loss functions for $(\mathbb{E}, \mathrm{var})$ are of this form; see Section 3.1 in Frongillo and Kash (2020).

2.3 Isotonic regression

2.3.1 General results

Suppose we have pairs of observations $(z_1, y_1), \dots, (z_n, y_n)$, where $y_1, \dots, y_n$ are real-valued, and the covariates $z_1, \dots, z_n$ are equipped with a total order, and $z_1 < z_2 < \cdots < z_n$. Repeated observations can easily be accommodated; see Remark 3.1 in Jordan et al. (2019). We aim to fit a function $\hat{g} = (\hat{g}_1, \hat{g}_2) : \{z_1, \dots, z_n\}^2 \to \mathbb{R}^2$ to these observations, such that g_1 is isotonic and models the conditional functional T given the covariates z_i, and g_2 is antitonic and models the conditional Bayes risk $\underline{L}$ given at (2.1) given the covariates z_i for some consistent loss function $L \in \mathcal{L}$. That is, if $z_i \le z_j$ then $\hat{g}_1(z_i) \le \hat{g}_1(z_j)$ and $\hat{g}_2(z_i) \ge \hat{g}_2(z_j)$, respectively. Considering the pair $(q_\alpha, \mathrm{ES}_\alpha)$ for example, one would be interested in an isotonic $\hat{g}_1$ and an antitonic $\hat{g}_2$ since $q_\alpha(Y_1) \le q_\alpha(Y_2)$ and $\mathrm{ES}_\alpha(Y_1) \ge \mathrm{ES}_\alpha(Y_2)$ whenever $Y_1 \le Y_2$ almost surely. Keeping this leading example in mind, we focus on the case that g_1 is isotonic, or increasing, and g_2 is decreasing, or antitonic. Adaptations of the results, where g_1 is desired to be decreasing or g_2 to be increasing are straight forward.

Following the literature on loss functions for expected shortfall, we first consider loss functions of the form (2.2) with $L' = 0$ (Nolde and Ziegel, 2017; Patton et al., 2019). When studying simultaneous optimality of solutions in Section 2.3.3, we also consider $L' \neq 0$. Let $h : \mathbb{R} \to \mathbb{R}$ positive and decreasing with $\lim_{x\to\infty} h(x) = 0$, $H(r) = \int_0^r h(x)\,\mathrm{d}x$. The goal is to minimize

$$\sum_{i=1}^{n} \tilde{L}(g_1(z_i), g_2(z_i), y_i) = \sum_{i=1}^{n} \big(H(g_2(z_i)) + h(g_2(z_i))(L(g_1(z_i), y_i) - g_2(z_i))\big) \tag{2.6}$$

over all functions $g = (g_1, g_2) : \{z_1, \dots, z_n\}^2 \to \mathbb{R}^2$ such that g_1 is increasing and g_2 is decreasing. Keeping either g_1 or g_2 fixed, we can give an optimal solution with respect to the other component.

Proposition 2.3.1. *(a) Let $g_1 : \{z_1, \dots, z_n\} \to \mathbb{R}$ be given. Then, the optimal antitonic solution $\hat{g}_2$ of (2.6) with g_1 fixed is given by*

$$\hat{g}_2(z_\ell) = -\min_{j\ge\ell} \max_{i\le j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i\le\ell} \min_{j\ge i} -\mathbb{E}(\bar{P}_{i:j}), \quad \ell = 1, \dots, n,$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(g_1(z_i), y_i), \dots, L(g_1(z_j), y_j)$.

(b) Let $g_2 : \{z_1, \dots, z_n\} \to \mathbb{R}$ be given. Then, any optimal isotonic solution $\hat{g}_1$ of (2.6) with g_2 fixed satisfies

$$\min_{j\ge\ell} \max_{i\le j} T^-(P^w_{i:j}) \le \hat{g}_1(z_\ell) \le \max_{i\le\ell} \min_{j\ge i} T^+(P^w_{i:j}),$$

where $P^w_{i:j}$ is the weighted empirical distribution of $y_i, \dots, y_j$ with weights proportional to $h(g_2(z_i)), \dots, h(g_2(z_j))$.

Proof. (a) Notice that for fixed g_1, the loss function (2.6) is a Bregman loss function. Moreover, $\hat{g}_2$ is isotonic if, and only if, $-\hat{g}_2$ is antitonic. Thus, we can solve the classical isotonic regression problem as in Jordan et al. (2019) for $-\hat{g}_2$ to obtain the optimal antitonic $\hat{g}_2$.

(b) Minimizing (2.6) for fixed g_2 is equivalent to minimizing

$$\sum_{i=1}^{n} h(g_2(z_i))L(g_1(z_i), y_i).$$

Using the same reasoning as in Remark 3.1 in Jordan et al. (2019), we have $h(g_2(z_i))L(g_1(z_i), y_i) = L(g_1(z_i), P^w_{i:i})$. Finally, Proposition 3.6 in Jordan et al. (2019) yields the result. □

If T is singleton-valued, Proposition 2.3.1 yields the existence and a necessary condition of a solution to (2.6).

Corollary 2.3.2. *If T is singleton-valued a solution $\hat{g}_1, \hat{g}_2$ to (2.6) exists. In particular, we have*

$$\hat{g}_2(z_\ell) = -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(\hat{g}_1(z_i), y_i), \dots, L(\hat{g}_1(z_j), y_j)$, and

$$\hat{g}_1(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} T(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T(P^w_{i:j}),$$

where $P^w_{i:j}$ is the weighted empirical distribution of $y_i, \dots, y_j$ with weights proportional to $h(\hat{g}_2(z_i)), \dots, h(\hat{g}_2(z_j))$.

Proof. For all solutions that are given by a min-max-representation with respect to some functional $\tilde{T}$ there exists a partition $\mathcal{Q}$ of the index set with $g(z_\ell) = \tilde{T}(Q)$, $\ell \in Q$, $Q \in \mathcal{Q}$ (Jordan et al., 2019, Proposition 4.17). Since there exist only finitely many partitions of the index set $\{1, \dots, n\}$ there exist only finitely many possible solutions. Therefore, an optimal solution has to exist. In particular, $\hat{g}_1$ has to be the solution obtained from Proposition 2.3.1 when $\hat{g}_2$ is treated as fixed and vice versa. Otherwise we could replace $\hat{g}_1$ by the solution obtained from Proposition 2.3.1 to obtain a smaller loss. Similarly, we could replace $\hat{g}_2$ by the solution in Proposition 2.3.1 to obtain a smaller loss. □

Furthermore, Proposition 2.3.1 suggests an algorithm for finding minimizers of (2.6), which roughly consists of the following steps:

1. Take g_2 constant and find the optimal $\hat{g}_1^{(1)}$.
2. Find the optimal $\hat{g}_2^{(1)}$ given $\hat{g}_1^{(1)}$.
3. Find the optimal $\hat{g}_1^{(2)}$ given $\hat{g}_2^{(1)}$.
4. Iterate steps 2 and 3 until $\hat{g}_1^{(k)} = \hat{g}_1^{(k-1)}$.

There is a problem with this algorithm if T is interval-valued, since then, the solution in part (b) of Proposition 2.3.1 is not unique. It turns out that it is best to choose the smallest possible solution corresponding to T^-, see Section 2.3.2 for details.

Fissler and Ziegel (2019) show that the expectation of consistent loss functions has no local minima. The optima in the isotonic regression case are more complex. But we believe that order sensitivity can be exploited to argue that the above algorithm can only converge to a global optimum. Numerical considerations where we perturbed the initial solutions to see whether they still converge to the same solution reinforced our suspicions that the algorithm does not converge to a saddle point. A rigorous mathematical proof for this conjecture, however, remains an open problem, for now.

2.3.2 Solution to the optimization problem

In this section, will show that for fixed g_2 it is best so choose

$$\hat{g}_1^-(z_\ell) := \min_{j \geq \ell} \max_{i \leq j} T^-(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P^w_{i:j}), \tag{2.7}$$

where P^w is the weighted empirical distribution with weights proportional to $h(\hat{g}_2(z_i)), \dots, h(\hat{g}_2(z_j))$, to minimize (2.6). For the investigations ahead we need to introduce some notation. We denote $T^\lambda = \lambda T^- + (1-\lambda)T^+$, $\lambda \in [0,1]$, where T^- and T^+ are the lower and upper bound of T, respectively. In (2.7) the indices ℓ, i and j are all elements of the index set $\{1, \dots, n\}$. If we were to restrict ℓ, i and j to be elements of the subset $\{1, \dots, m\}$, $m \leq n$, we would obtain an optimal solution on the subset $(z_1, y_1), \dots, (z_m, y_m)$ of the original data set. In the following, we denote an optimal solution on this subset by $\hat{g}_{1;1:m}$ and by $\hat{g}_1|_{1:m}$ we denote the optimal solution on the original set restricted to $\{z_1, \dots, z_m\}$.

The following auxiliary result relates $\hat{g}_{1;1:m}$ to $\hat{g}_1|_{1:m}$ in the case where $\hat{g}_1$ is given by a min-max-representation.

Lemma 2.3.3. *Assume that*

$$\hat{g}_1(z_\ell) := \min_{j \geq \ell} \max_{i \leq j} T^\lambda(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^\lambda(P^w_{i:j})$$

for some $\lambda \in [0,1]$. *Then we have* $\hat{g}_1|_{1:m} \leq \hat{g}_{1;1:m}$.

Proof. Notice that

$$\hat{g}_{1;1:m}(z_\ell) = \min_{\substack{j \geq \ell \\ j \leq m}} \max_{i \leq j} T^\lambda(P^w_{i:j}) \geq \min_{j \geq \ell} \max_{i \leq j} T^\lambda(P^w_{i:j}) = \hat{g}_1(z_\ell). \qquad \square$$

Before we continue, let us recall some observations made in Jordan et al. (2019). For fixed weights, we can minimize

$$\sum_{i=1}^{n} \mathbb{1}\{\eta \leq \hat{g}_1(z_i)\} V(\eta, P^w_{i:i}), \quad \text{for all } \eta \in \mathbb{R}$$

to obtain a solution to (2.6). Because we want $\hat{g}_1$ to be isotonic, this means that for a given $\eta \in \mathbb{R}$ we have to find an index $\ell \in \{1. \ldots, n+1\}$ that minimizes

$$\sum_{i=\ell}^{n} V(\eta, P^w_{i:i}). \tag{2.8}$$

The search for the optimal index ℓ needs to be conducted for every $\eta \in \mathbb{R}$. For $\eta \in \mathbb{R}$ we denote the set of indices minimizing (2.8) by $I_{1:n}(\eta)$.

Recall that optimal solutions $\hat{g}_1$ are in one-to-one correspondence to increasing, left-continuous functions $\iota : \mathbb{R} \to \{1, \ldots, n+1\}$ with $\iota(\eta) \in I_{1:n}(\eta)$, for all $\eta \in \mathbb{R}$, in the sense that

$$\inf\{\eta : \iota(\eta) > \ell\} = \hat{g}_1(z_\ell) = \max\{\eta : \iota(\eta) \leq \ell\}.$$

Thus, any solution to the isotonic regression problem imposes a minimizing index $\iota(\eta)$ for every $\eta \in \mathbb{R}$.

The next result shows that if $\hat{g}_1$ is a solution to the isotonic regression problem (2.6) with $\hat{g}_1(z_m) < \hat{g}_1(z_{m+1})$ then $\hat{g}_1|_{1:m}$ is an optimal solution to the isotonic regression problem (2.6) on the subsample $(z_1, y_1), \ldots, (z_m, y_m)$.

Lemma 2.3.4. *We have that $I_{1:n}(\eta) \cap \{1, \ldots, m+1\} \subseteq I_{1:m}(\eta)$, where $I_{1:m}(\eta)$ is the set of minimizing indices for the isotonic regression problem (2.6) on the subsample $(z_1, y_1), \ldots, (z_m, y_m)$.*

Proof. Let $\ell \in I_{1:n}(\eta) \cap \{1, \ldots, m\}$ for some $\eta \in \mathbb{R}$. Therefore, the function

$$t_\eta : \{1, \ldots, n+1\} \to \mathbb{R}, x \mapsto \sum_{i=x}^{n} V(\eta, P^w_{i:i})$$

has a minimum at ℓ. We can write

$$\sum_{i=\ell}^{n} V(\eta, P_{i:i}^w) = \sum_{i=\ell}^{m} V(\eta, P_{i:i}^w) + \sum_{i=m+1}^{n} V(\eta, P_{i:i}^w).$$

Hence, $t_\eta|_{1:m}$ has also a minimum at ℓ and thus $\ell \in I_{1:m}(\eta)$. If t_η has a minimum at $\ell = m+1$ then

$$t_\eta(x) - \sum_{i=m+1}^{n} V(\eta, P_{i:i}^w) \geq 0$$

with equality for $x = m+1$. Thus, $I_{1:n}(\eta) \cap \{1, \ldots, m+1\} \subseteq I_{1:m}(\eta)$. □

Corollary 2.3.5. *Let $\hat{g}_1$ be a solution* (2.6) *with $\hat{g}_1(z_m) < \hat{g}_1(z_{m+1})$. Then we have that $\hat{g}_1|_{1:m}$ is an optimal solution to* (2.6) *on the subsample $(z_1, y_1), \ldots, (z_m, y_m)$.*

We now would like to show that for fixed weights the solution

$$\hat{g}_1^-(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} T^-(P_{i:j}^w) = \max_{i \leq \ell} \min_{j \geq i} T^-(P_{i:j}^w)$$

is most likely to minimize (2.6). An intuition behind this statement is obtained by combining Lemma 2.3.3 with Lemma 3.4 from Jordan et al. (2019). The statement in Lemma 2.3.3 is equivalent to $\hat{g}_1^-|_{(m+1):n} \geq \hat{g}_{1,(m+1):n}^-$. Lemma 3.4 of Jordan et al. (2019) on the other hand, implies that any optimal solution $\hat{g}_{1,(m+1):n}$ on $(z_{m+1}, y_{m+1}), \ldots, (z_n, y_n)$ has to satisfy $\hat{g}_{1,(m+1):n}^- \leq \hat{g}_{1,(m+1):n} \leq \hat{g}_{1,(m+1):n}^+$. Thus, $\hat{g}_1^-$ has the highest chance to lie between those bounds.

To prove this formally the *order sensitivity* of loss functions is needed. We recall the definition given in Steinwart et al. (2014).

Definition 2.3.6. Let $\mathcal{P}$ be a class of probability distributions. A loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is said to be *$\mathcal{P}$-order sensitive for T*, if the image of T is an interval, and for all $P \in \mathcal{P}$ and all $t_1, t_2 \in \mathbb{R}$ with either $t_2 < t_1 \leq T^-(P)$ or $T^+(P) \leq t_1 < t_2$, we have $L(t_1, P) < L(t_2, P)$.

It follows directly from the definition that order sensitive loss functions are consistent. The reverse holds under weak regularity conditions on the functional; see Lambert (2019, Proposition 11). The loss functions in class $\mathcal{L}$ are order-sensitive because they are defined via oriented identification function and a positive measure H_1 (Steinwart et al., 2014, Theorem 7). Thus, the loss function L in the following proposition is order sensitive.

Proposition 2.3.7. *For fixed g_2, $\hat{g}_1^-$ given by* (2.7) *and any increasing $\hat{g}_1$ we have*

$$\sum_{i=1}^{n} \tilde{L}(\hat{g}_1^-(z_i), g_2(z_i), y_i) \leq \sum_{i=1}^{n} \tilde{L}(\hat{g}_1(z_i), g_2(z_i), y_i).$$

Proof. Note that for each $\hat{g}_1$ and we have a partition $\mathcal{Q}$ of the index set such that

$$\hat{g}_1(z_i) = \hat{g}_1(z_j) \quad \text{for all } i, j \in Q,\ Q \in \mathcal{Q}.$$

We let Q_m denote the partition element corresponding to $\hat{g}_1$ containing m, and Q_m^- denote the partition element corresponding to $\hat{g}_1^-$ containing m.

By Lemma 2.2.1, it suffices to show that for all $\eta \in \mathbb{R}$

$$\sum_{i=1}^{n} S_{\eta,2}(\hat{g}_1^-(z_i), g_2(z_i), y_i) \leq \sum_{i=1}^{n} S_{\eta,2}(\hat{g}_1(z_i), g_2(z_i), y_i).$$

For the latter, it suffices to show that for all $m \leq n$

$$\sum_{\ell=m}^{n} L(\hat{g}_1^-(z_\ell), y_\ell) \leq \sum_{\ell=m}^{n} L(\hat{g}_1(z_\ell), y_\ell). \tag{2.9}$$

This statement clearly holds if $\hat{g}_1$ has a jump in a some non-minimizing index, that is ℓ with $\ell \notin \cup_\eta I_{1:n}(\eta)$. Thus, we can focus on $\hat{g}_1$ that solely jumps in ℓ with $\ell \in \cup_\eta I_{1:n}(\eta)$. This implies that we have

$$\sum_{\ell=1}^{n} L(\hat{g}_1^-(z_\ell), y_\ell) = \sum_{\ell=1}^{n} L(\hat{g}_1(z_\ell), y_\ell).$$

In the following, we will prove the converse to (2.9), that is, for all $m \leq n$ we have

$$\sum_{\ell=1}^{m} L(\hat{g}_1(z_\ell), y_\ell) \leq \sum_{\ell=1}^{m} L(\hat{g}_1^-(z_\ell), y_\ell). \tag{2.10}$$

If $m = \max Q_m$, it follows from Corollary 2.3.5 that $\hat{g}_1|_{1:m}$ is optimal on $(z_1, y_1), \ldots, (z_m, y_m)$ and therefore (2.10) holds.

For $m \neq \max Q_m$, we distinguish two cases.

Case 1: If $m = \max Q_m^-$, it follows from Lemma 2.3.3 and Proposition 2.3.1 that

$$\hat{g}_1^-|_{1:m} = \hat{g}_{1;1:m}^- \leq \hat{g}_1|_{1:m} \leq \hat{g}_1^+|_{1:m} \leq \hat{g}_{1;1:m}^+.$$

By Lemma 2.3.4 we have $I_{1:n}(\eta) \cap \{1, \ldots, m+1\} \subseteq I_{1:m}(\eta)$ for all $\eta \in \mathbb{R}$. Hence, $\iota|_{1:m}(\eta) \in I_{1:m}(\eta)$ for all $\eta \in \mathbb{R}$, where $\iota : \mathbb{R} \to \{1, \ldots, n+1\}$ is the function imposing the score minimizing-indices corresponding to $\hat{g}_1$. Thus, Proposition 3.5 in Jordan et al. (2019) implies that $\hat{g}_1|_{1:m}$ is a solution to the isotonic regression problem on $(z_1, y_1), \ldots, (z_m, y_m)$.

Case 2: Consider the case $m \neq \max Q_m$ and let $j = \max(\min Q_m, \min Q_m^-)$. It follows from the previous considerations that $\hat{g}_1$ is optimal up to $j - 1$ in the sense that it is a

minimizer on $(z_1, y_1), \ldots, (z_{j-1}, y_{j-1})$. We know that $\hat{g}_1^-|_{1:m} \leq \hat{g}_{1;1:m}^-$ so if $\hat{g}_1|_{1:m} \geq \hat{g}_{1;1:m}^-$ we can conclude with the same reasoning as in case 1.

Otherwise, let $j_0 \geq j$ be the minimal index with $\hat{g}_{1;1:m}^-(z_{j_0}) > \hat{g}_1|_{1:m}(z_{j_0})$. Clearly $j_0 \in Q_m$ and hence $\hat{g}_1$ is constant on $\{j, \ldots, j_0\}$. Moreover, if $j_0 > j$ then for all $\ell \in \{1, \ldots, j_0 - 1\}$ we have that

$$\hat{g}_{1;1:m}^-(z_\ell) = \hat{g}_{1;1:(j_0-1)}^-(z_\ell) \leq \hat{g}_1(z_\ell) \leq \hat{g}_1^+|_{1:(j_0-1)}(z_\ell) \leq \hat{g}_{1;1:(j_0-1)}^+(z_\ell),$$

implying that $\hat{g}_1$ is in fact optimal up to $j_0 - 1$. Of course, if $j_0 = j$, we already know that $\hat{g}_1$ is optimal up to $j_0 - 1$, since we know that $\hat{g}_1$ is optimal up to $j - 1$ from our previous considerations. Thus, it remains to check what happens for $\ell \in \{j_0, \ldots, m\}$.

For $\ell \in \{j_0, \ldots, m\}$ we have $\hat{g}_1^-(z_\ell) = c^- \leq c = \hat{g}_1(z_\ell) < \hat{g}_{1;1:m}^-(z_\ell)$ for some constants c^- and c.

Denote by $Q_{s;1:m}^-, \ldots, Q_{r;1:m}^-$ the partition elements of $\hat{g}_{1;1:m}^-$ on $\{j_0, \ldots, m\}$. Then, for $k \in \{s, \ldots, r\}$ we have

$$\sum_{\ell \in Q_{k;1:m}^-} L(\hat{g}_{1;1:m}^-, y_\ell) \leq \sum_{\ell \in Q_{k;1:m}^-} L(c, y_\ell) \leq \sum_{\ell \in Q_{k;1:m}^-} L(c^-, y_\ell)$$

since $\hat{g}_{1;1:m}^-$ is constant each $Q_{k;1:m}^-$ and L is order-sensitive. Therefore, (2.10) is fulfilled. □

Finally, we have all necessary results to see that $\hat{g}_1^-$ is indeed our best bet. Define

$$\hat{g}_1^-(z_\ell; w) := \min_{j \geq \ell} \max_{i \leq j} T^-(P_{i:j}^w) = \max_{i \leq \ell} \min_{j \geq i} T^-(P_{i:j}^w)$$
$$\hat{g}_2^-(z_\ell; \hat{g}_1^-) := -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(\hat{g}_1^-(z_i), y_i), \ldots, L(\hat{g}_1^-(z_j), y_j)$ and $P_{i:j}^w$ is the weighted empirical distribution of $y_i, \ldots, y_j$ with weights w.

Proposition 2.3.8. *Assume that there exist $\hat{g}_1, \hat{g}_2 \colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ minimizing (2.6), then $\hat{g}_1^-(\cdot; h(\hat{g}_2)), \hat{g}_2^-(\cdot; \hat{g}_1^-(\cdot; h(\hat{g}_2)))$ are also minimizers.*

Proof. Clearly the pair $\hat{g}_1(\cdot), \hat{g}_2(\cdot)$ has to satisfy the restrictions imposed by Proposition 2.3.1 as otherwise they would not be optimal. Proposition 2.3.7 implies that the pair $\hat{g}_1^-(\cdot; h(\hat{g}_2)), \hat{g}_2(\cdot)$ is also a minimizing pair to (2.6). Finally applying part (a) of Proposition 2.3.7 we can conclude that $\hat{g}_2(\cdot) = \hat{g}_2^-(\cdot; \hat{g}_1^-(\cdot; h(\hat{g}_2)))$. □

2.3.3 Simultaneously optimal solutions

In Section 2.3.2 we solved the isotonic regression problem (2.6). To consider simultaneous optimality we let $L' \neq 0$. A simultaneously optimal solution $\hat{g}_1, \hat{g}_2$ therefore has to minimize the expected elementary losses

$$\sum_{i=1}^{n} S_{\eta,1}(g_1(z_i), y_i) \tag{2.11}$$

and

$$\sum_{i=1}^{n} S_{\eta,2}(g_1(z_i), g_2(z_i), y_i) \tag{2.12}$$

for all $\eta \in \mathbb{R}$ among all increasing functions $g_1 : \{z_1, \dots, z_n\} \to \mathbb{R}$ and all decreasing functions $g_2 : \{z_1, \dots, z_n\} \to \mathbb{R}$. The expected elementary score (2.11) is minimized if and only if $\hat{g}_1$ is optimal isotonic solution with respect to T characterized in Jordan et al. (2019). Thus, there can only exist a simultaneously optimal solution if for one such $\hat{g}_1$ there exists $\hat{g}_2 : \{z_1, \dots, z_n\} \to \mathbb{R}$ decreasing so that the pair $\hat{g}_1, \hat{g}_2$ minimizes (2.12) for all $\eta \in \mathbb{R}$.

The proof of Proposition 2.3.7 suggests that for any $m \leq n$

$$\sum_{i=m}^{n} L(\hat{g}_1^-(z_i), y_i) \leq \sum_{i=m}^{n} L(\hat{g}_1(z_i), y_i)$$

with equality whenever $m = n$. Note that minimizing (2.12) for all $\eta \in \mathbb{R}$ is equivalent to minimizing

$$\sum_{i=1}^{n} \mathbb{1}\{\eta \leq -g_2(z_i)\} L(g_1(z_i), y).$$

Thus, a pair $\hat{g}_1, \hat{g}_2$ can only be simultaneously optimal if $\hat{g}_1|_{m:n}$ is an optimal isotonic solution on $(z_m, y_m), \dots, (z_n, y_n)$ for all $m \in \{1, \dots, n\}$ with $\hat{g}_2(z_{m-1}) > \hat{g}_2(z_m)$. If this is not the case for some $m \in \{1, \dots, n\}$, we can find $\hat{g}_1^m$ such that the pair $\hat{g}_1^m, \hat{g}_2^m$, where $\hat{g}_2^m$ is the corresponding solution obtained via Proposition 2.3.1, dominates $\hat{g}_1, \hat{g}_2$ for all $\eta \in \mathbb{R}$ with $\hat{g}_2(z_{m-1}) < \eta \leq \hat{g}_2(z_{m-1})$. But inevitably this solution performs worse for other $\eta \in \mathbb{R}$, especially for $\eta \leq -g_2(z_1)$. Figure 2.1 displays a data example where a simultaneously optimal solution does not exist because there exists some index m with $\hat{g}_2(z_{m-1}) > \hat{g}_2(z_m)$ but $\hat{g}_1^-|_{m:n}$ is not an optimal isotonic solution on $(z_m, y_m), \dots, (z_n, y_n)$. The previous considerations are captured by the following proposition.

Proposition 2.3.9. *A simultaneously optimal solution exists if and only if $\hat{g}_1^-|_{m:n}$ is an optimal solution on $(z_m, y_m), \dots, (z_n, y_n)$ for all $m \in \{2, \dots, n\}$ such that $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ and $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$.*

Proposition 2.3.9 supplies us with a relatively straight forward criterion to check for simultaneous optimality. The approach is to first calculate

$$\hat{g}_1^-(z_\ell) := \min_{j\geq\ell}\max_{i\leq j} T^-(P_{i:j}) = \max_{i\leq\ell}\min_{j\geq i} T^-(P_{i:j}),$$
$$\hat{g}_2^-(z_\ell) := -\min_{j\geq\ell}\max_{i\leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i\leq\ell}\min_{j\geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

with $\bar{P}$ as defined in Proposition 2.3.1. In a second step, for each $m \geq 2$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ and $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$ one has check whether $\hat{g}_1^-|_{m:n}$ is an optimal solution on the subset $(z_m, y_m), \ldots, (z_n, y_n)$. To check whether $\hat{g}_1^-|_{m:n}$ remains optimal we can compare the expected elementary score for $\hat{g}_1^-|_{m:n}$ to the one of $\hat{g}_{1;m:n}^-$. If $\hat{g}_1^-|_{m:n}$ remains optimal for each $m \geq 2$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ and $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$, then the solution $(\hat{g}_1^-, \hat{g}_2^-)$ is indeed simultaneously optimal.

For bivariate functionals T with two elicitable components there always exists a subclass of loss functions $L(x_1, x_2, y)$ admitting a mixture representation that can be separated into two parts, the first only depending on x_1 and the second only depending on x_2. Namely

$$\mathcal{L}_2 = \Bigg\{ \int_{\mathbb{R}} S_{\eta,1}(x_1, y)\,\mathrm{d}H_1(\eta) + \int_{\mathbb{R}} S_{\eta,2}(x_2, y)\,\mathrm{d}H_2(\eta) : \\ H_1, H_2 \text{ non-negative measures on } \mathbb{R} \Bigg\},$$

where

$$S_{\eta,1}(x_1, y) = (\mathbb{1}\{\eta \leq x_1\} - \mathbb{1}\{\eta \leq y\})\, V_1(\eta, y)$$
$$S_{\eta,2}(x_2, y) = (\mathbb{1}\{\eta \leq x_2\} - \mathbb{1}\{\eta \leq y\})\, V_2(\eta, y)$$

with V_1 and V_2 being the respective identification functions of the components. Thereby minimizing $\mathbb{E}L(g_1(Z), g_2(Z), Y)$, simultaneously over $L \in \mathcal{L}_2$, among all increasing g_1 and decreasing g_2 can be split into two independent optimization problems. In this case Jordan et al. (2019) provide all necessary tools for a complete characterization of all solutions. But not all consistent loss functions lie necessarily in $\mathcal{L}_2$. If T is a vector of moments this can be seen in Proposition 4.11 in Fissler and Ziegel (2019). In the case where T is a vector of quantiles, however, $\mathcal{L}_2$ comprises all consistent losses (Fissler and Ziegel, 2016, Proposition 4.2) explaining some of the optimality properties of the IDR introduced by Henzi et al. (2019). Thus, when considering functionals with elicitable components one can reach simultaneous optimality at least with respect to the class $\mathcal{L}_2$. When considering functionals with elicitation complexity greater than one however, the elementary loss for the Bayes risk always depends on the first component, so that possibly no simultaneous optimum exists.

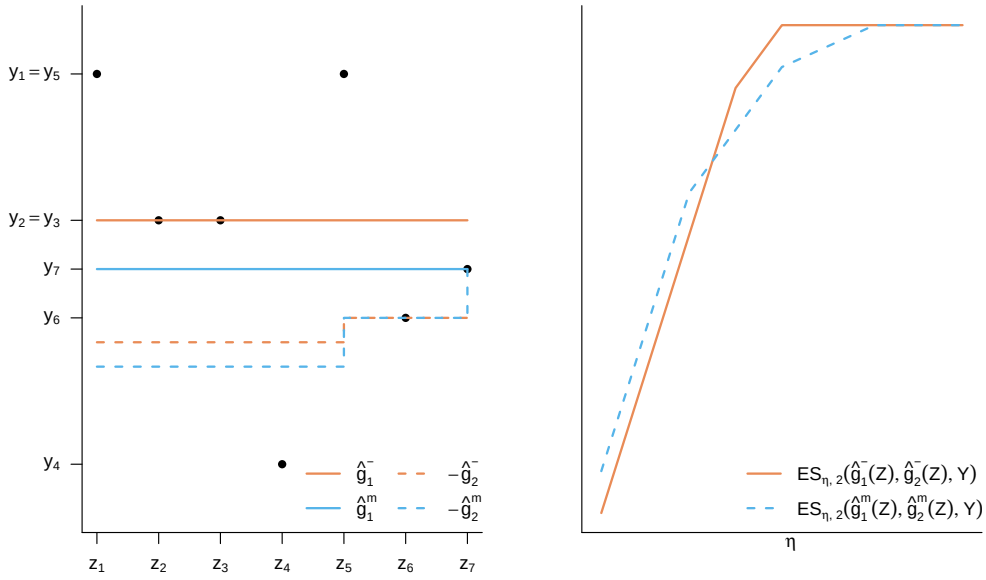

Figure 2.1: Consider the pair $\underline{T} = (q_{0.5}, \mathrm{ES}_{0.5})$. Then, for this specific sample of seven data points (black) on the left, we have that $\hat{g}_1^-|_{5:7}$ is not an optimal isotonic solution on (z_5, y_5), (z_6, y_6), (z_7, y_7) but $\hat{g}_2^-(z_4) > \hat{g}_2^-(z_5)$. The function $\hat{g}_1^m$, on the other hand, is clearly not an optimal isotonic solution to the global optimization problem. However, the Murphy diagram on the right shows the there are indeed values of η where the pair $\hat{g}_1^m$, $\hat{g}_2^m$ has a smaller expected loss than the pair $\hat{g}_1^-$, $\hat{g}_2^-$.

2.4 Simulation study

We let

$$\hat{g}_1^-(z_\ell) := \min_{j \geq \ell} \max_{i \leq j} T^-(P_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P_{i:j})$$
$$\hat{g}_2^-(z_\ell) := -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j})$$

In the previous section, we have seen this is not necessarily a simultaneously optimal solution whenever the elicitation complexity is greater than one. However, we were able to formulate a criterion allowing us to check whether a solution is simultaneously optimal. This section is devoted to the investigation on how often simultaneous optimality occurs and the number of iterations needed to obtain an optimal solution for a specific loss function instead, whenever the solution $\hat{g}_1^-, \hat{g}_2^-$ is not simultaneously optimal. To this end we consider the two prominent examples (q_α, ES) and $(\mathbb{E}, \mathrm{var})$ in a simulation study.

First, let us examine what we would expect to result from those simulations in terms of simultaneous optimality. In Section 2.3.2, we saw that simultaneous optimality is attained whenever $\hat{g}_1^-|_{m:n}$ remains an optimal solution for all $m \in \{2, \ldots, n\}$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$. Clearly, this requirement is fulfilled as long as $\hat{g}_2^-$ jumps at the

same point as $\hat{g}_1^-$. Naturally, the more jumps $\hat{g}_1^-$ has, or equivalently the less pooling was required, the higher are the chances for simultaneous optimality, in that there are not many additional restrictions left to be imposed by $\hat{g}_2^-$. Thus, the less the isotonicity constraint is violated in the data the higher the chances for the pair $(\hat{g}_1^-, \hat{g}_2^-)$ to be simultaneously optimal. Only considering the impact of $\hat{g}_1^-$, we would expect the chance for simultaneous optimality to decrease with increasing variance in the data. Moreover, for fixed variance we would expect the chance of simultaneous optimality to decrease with increasing sample size, because the chance for necessary pooling increases.

Now let us think about the impact of $\hat{g}_2^-$. We have seen in Proposition 2.3.1 that $\hat{g}_2^-$ is fitted to the transformed data points $(z_1, L(\hat{g}_1^-(z_1), y_1)), \dots, (z_n, L(\hat{g}_1^-(z_n), y_n))$, where the transformed y-values depend on the loss L of y_ℓ and $\hat{g}_1^-(z_\ell)$. The order sensitivity of the loss function ensures that the transformation $L(\hat{g}_1^-(z_\ell), y_\ell)$ takes larger values when $\hat{g}_1^-(z_\ell)$ and y_ℓ are far apart and smaller values when they are close. Thus, if small modifications are necessary to obtain $\hat{g}_1$, then we would expect the transformed data to be approximately constant. The outcome of the transformation however depends on how the loss L weighs the differences.

Let us check whether the simulations corroborate our thoughts. The setup for the simulations study was the following: For the pair $(q_\alpha, \mathrm{ES}_\alpha)$ we aimed to fit an increasing function $\hat{g}_1^-$ and decreasing function $\hat{g}_2^-$ to simulated data sets. In order to do so, we drew n points z_ℓ independently and uniformly from $[0, 100]$. The corresponding y-value was $y_\ell = z_\ell + \epsilon_\ell$ where $\epsilon_\ell \sim \mathcal{N}(0, \sigma^2)$ are independent and independent of z_ℓ. We let $n \in \{10, 100, 500, 1000\}$ and $\sigma \in \{3, 10, 20, 30\}$ and then we repeated the experiment $M = 1000$ times to count the number of times simultaneous optimality occurred. For each combination of sample size n and standard deviation σ, Figure 2.2 shows one of the generated data sets. To investigate whether the results differ depending on level α, we calculated $\hat{g}_1^-$ and $\hat{g}_2^-$ for each data set for all $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For a specific data set, Figure 2.3 shows the fits $\hat{g}_1^-$ for all levels α, and Figure 2.4 contains the corresponding fits $-\hat{g}_2^-$.

Sometimes it is more reasonable to assume that both g_1 and g_2 are increasing. All results in this manuscript can naturally be adapted to this case. For the pair $(\mathbb{E}, \mathrm{var})$ we therefore aimed to fit two increasing functions $\hat{g}_1^-$ and $\hat{g}_2^-$ to simulated data sets. Thus, again drew n points z independently and uniformly from $[0, 100]$. The corresponding y-value was $y_\ell = z_\ell + \epsilon_\ell$ where $\epsilon_\ell \sim \mathcal{N}(0, c\ell/\sqrt{n})$ were independent. We let $n \in \{10, 100, 500, 1000\}$ and $c \in \{0.5, 1, 3, 6\}$ and then generated $M = 1000$ data sets and calculated the corresponding fits $\hat{g}_1^-$ and $\hat{g}_2^-$. Figure 2.5 shows a generated data set for each combination of sample size n and constant c. Figure 2.6 contains the fits $\hat{g}_1^-$ and $\hat{g}_2^-$ for a specific data set.

Simultaneous optimality. Using the criterion in Proposition 2.3.9, we counted the number times simultaneous optimality occurred. Table 2.1 contains the results for the

pair $(q_\alpha, \mathrm{ES}_\alpha)$. The percentage of times simultaneous optimality is reached is displayed. It can be seen that the results confirm our expectations. With increasing sample size and increasing variance the percentage decreases drastically. The reason that not all levels α are equally affected is due to the different weights that L imposes depending on the level α.

The results for the pair $(\mathbb{E}, \mathrm{var})$ in Table 2.2 also confirm our expectations. The reason why the percentage in this case decreases even more rapidly is that the expectation $\mathbb{E}$ is less robust when it comes to removing data from a partition element than the quantile q_α is.

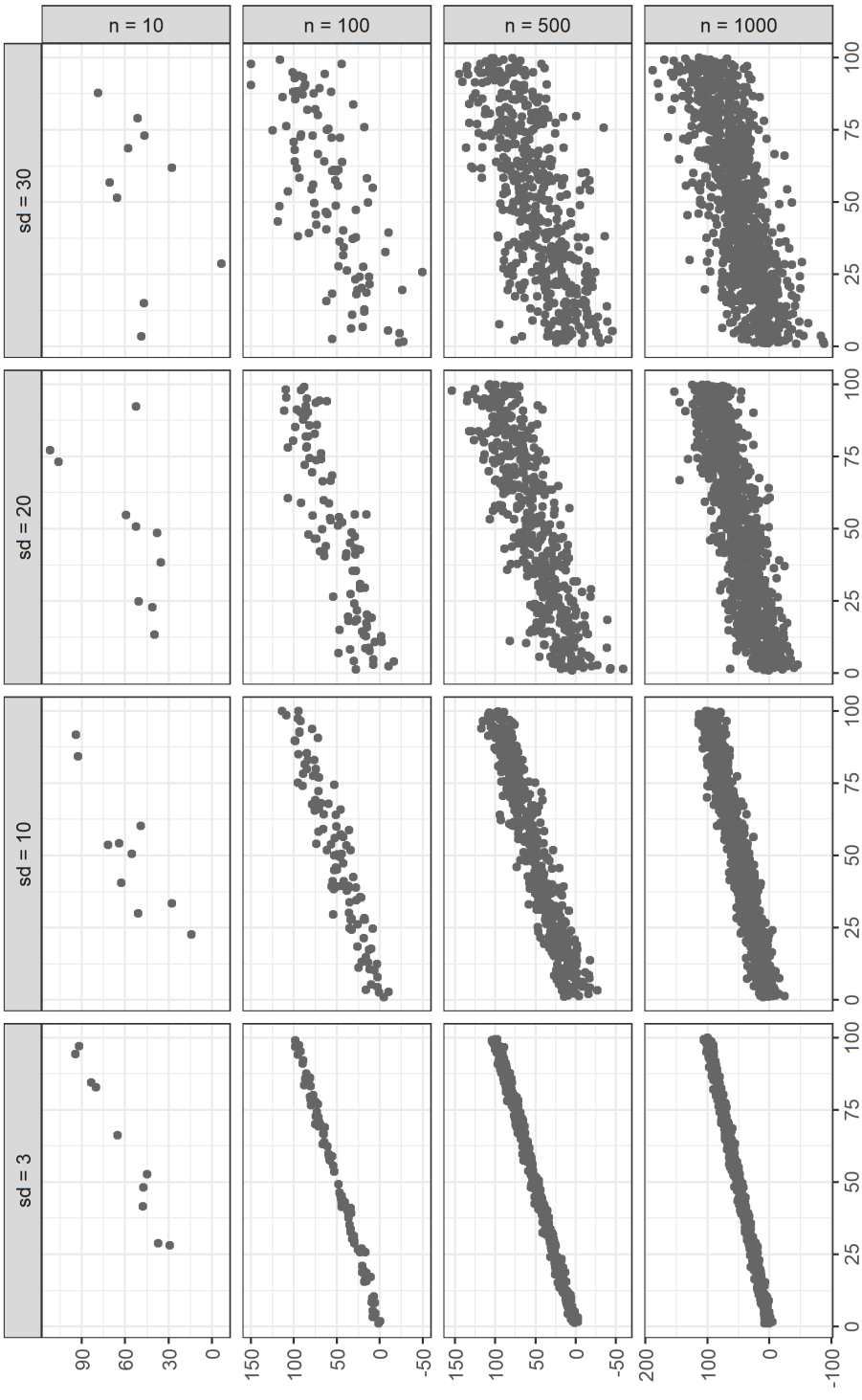

Figure 2.2: To get an idea of the data sets considered for the pair $(q_\alpha, \mathrm{ES}_\alpha)$ a specific data set is drawn for each combination of standard deviation σ and sample size n.

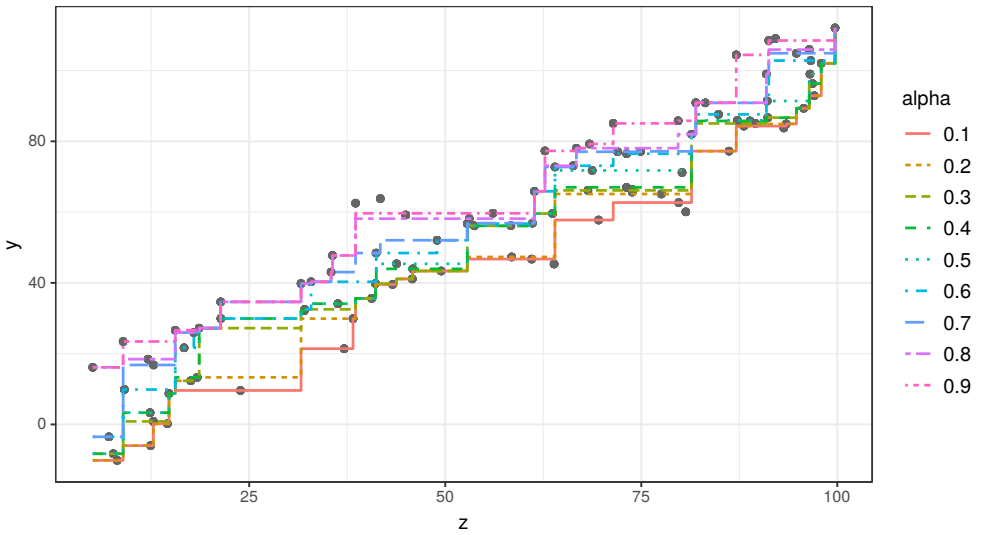

Figure 2.3: For a set of $n = 100$ data points and the pair $(q_\alpha, \mathrm{ES}_\alpha)$ the optimal fit $\hat{g}_1^-$ was drawn for each $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

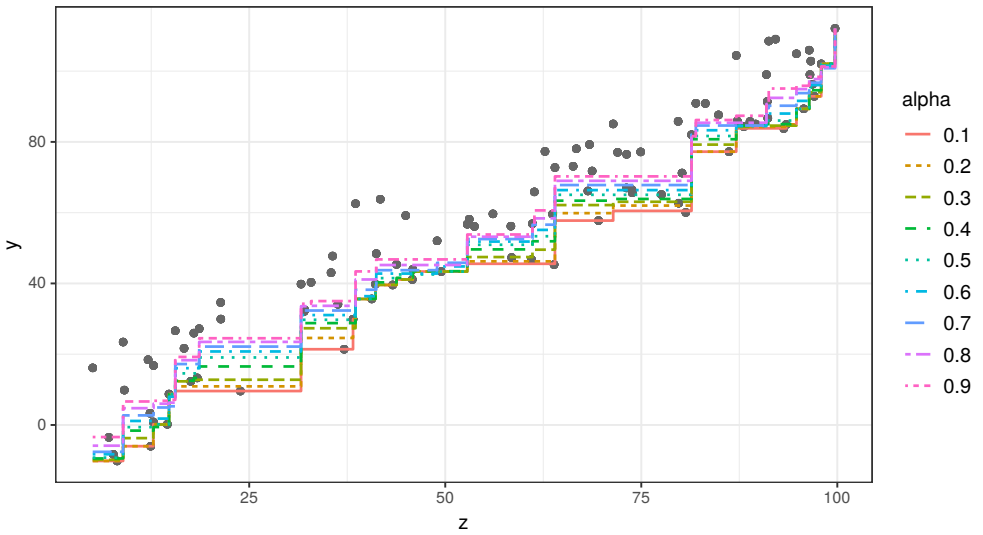

Figure 2.4: For the same choice of $n = 100$ data points as in Figure 2.3 the corresponding fits $\hat{g}_2^-$ are calculated and $-\hat{g}_2^-$ is displayed for each $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

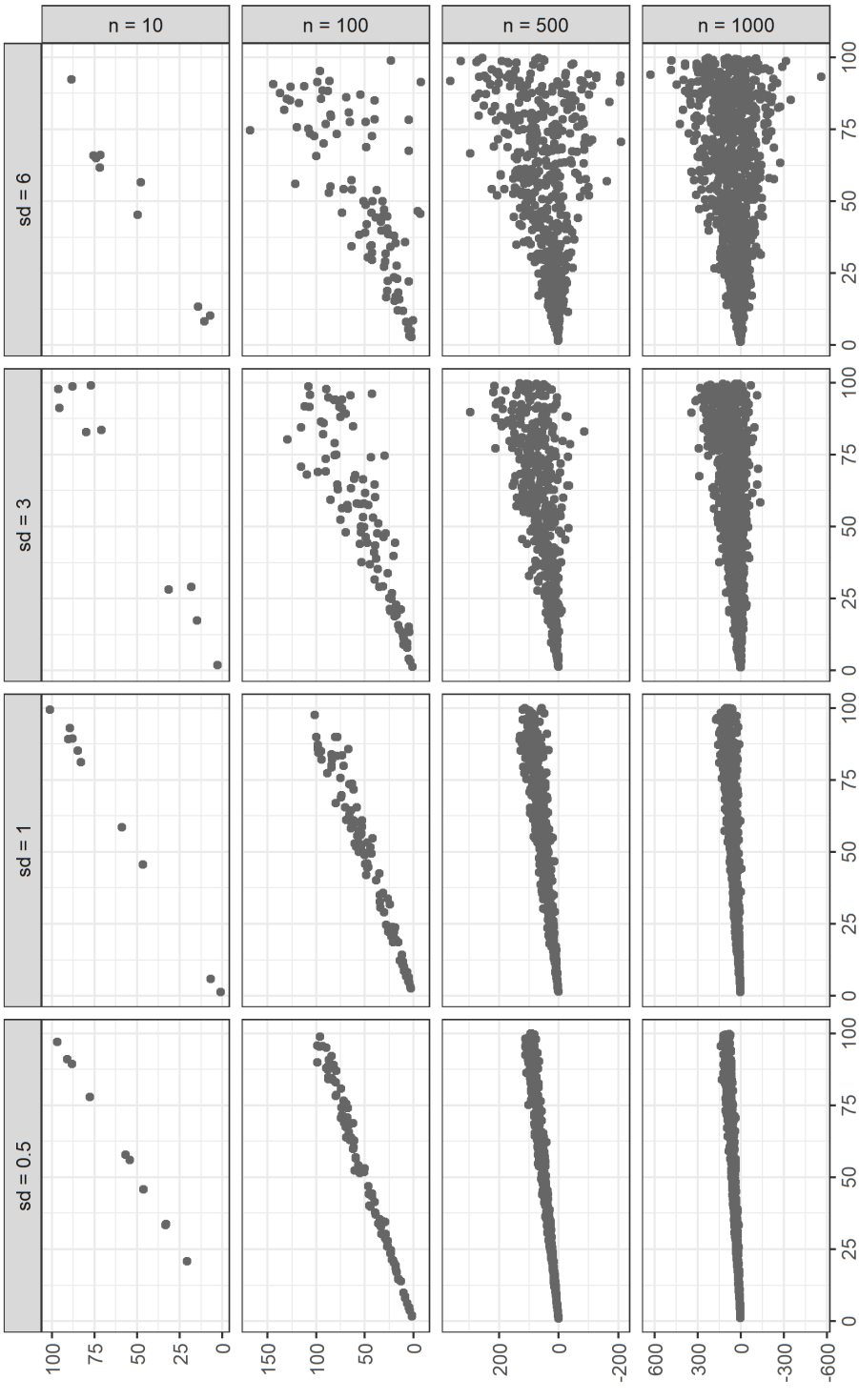

Figure 2.5: To get a feeling for the data sets considered for the pair $(\mathbb{E}, \text{var})$ a specific data set is drawn for each combination of constant c and sample size n.

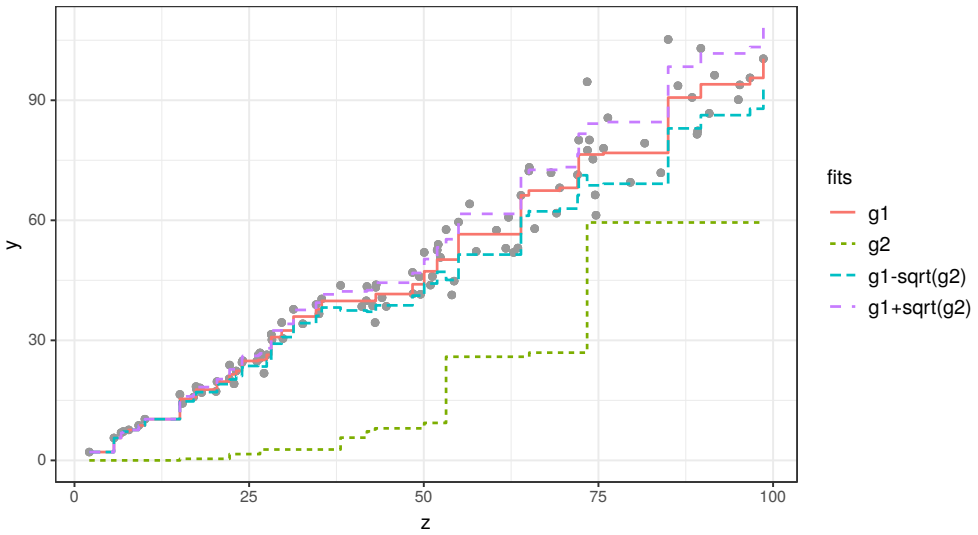

Figure 2.6: For a sample of $n = 100$ data points and the pair $(\mathbb{E}, \mathrm{var})$ the optimal fit $\hat{g}_1$ is drawn in red and $\hat{g}_2$ is in green. Moreover, $\hat{g}_1 - \sqrt{\hat{g}_2}$ and $\hat{g}_1 + \sqrt{\hat{g}_2}$ are drawn in blue and pink, respectively

Table 2.1: The percentage of times simultaneous optimality occurred for the pair $(q_\alpha, \mathrm{ES}_\alpha)$ for each combination of sample size n and standard deviation σ and level α is printed.

		$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
$n = 10$	$\sigma = 3$	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.94	0.94
	$\sigma = 10$	1.00	1.00	1.00	1.00	0.98	0.96	0.79	0.70	0.69
	$\sigma = 20$	1.00	1.00	1.00	0.98	0.94	0.92	0.69	0.58	0.50
	$\sigma = 30$	1.00	1.00	0.98	0.97	0.88	0.88	0.64	0.53	0.44
$n = 100$	$\sigma = 3$	1.00	0.97	0.60	0.48	0.14	0.13	0.00	0.00	0.00
	$\sigma = 10$	0.96	0.53	0.16	0.08	0.01	0.01	0.00	0.00	0.00
	$\sigma = 20$	0.80	0.31	0.11	0.06	0.01	0.02	0.00	0.00	0.00
	$\sigma = 30$	0.70	0.27	0.12	0.06	0.02	0.02	0.00	0.00	0.00
$n = 500$	$\sigma = 3$	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 1000$	$\sigma = 3$	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.2: The percentage of times simultaneous optimality occurred for the pair $(\mathbb{E}, \text{var})$ for each combination of sample size n and constant c is printed.

	$c = 0.5$	$c = 1$	$c = 3$	$c = 6$
$n = 10$	0.98	0.95	0.79	0.57
$n = 100$	0.02	0.00	0.00	0.00
$n = 500$	0.00	0.00	0.00	0.00
$n = 1000$	0.00	0.00	0.00	0.00

Specific loss function. We have previously seen that simultaneous optimality is often not attainable. In these cases, we have to choose a specific loss function to solve the isotonic regression problem. It is natural to ask, how different these solutions are compared to our candidate for simultaneous optimality. We address this question in the second part of this simulation study. For both examples, we choose two different weight function h and count the number of iterations the algorithm needed to get from the candidate for simultaneous optimality to the optimal solution for the specific loss. For the pair $(q_\alpha, \text{ES}_\alpha)$, we wanted to consider a loss $\tilde{L}$ used for the pair in applications. To this end, we consider the $(1/2)$-homogeneous loss from Nolde and Ziegel (2017). Because they use a different sign convention, we modified our method to match their sign convention. To this end we let $L(x, y) = -1/\alpha \mathbb{1}\{x < y\}(x - y) + x$ and fit an isotonic function $\hat{g}_2$ to the transformed data. The $(1/2)$-homogeneous from Nolde and Ziegel (2017) arises when choosing $h_1(x) = 1/(2\sqrt{x})$. To see the behavior with respect to a second set of weights, we additionally chose $h_2(x) = \exp(-x)$. The iteration was stopped when the loss given by (2.6) did not improve by more that 10^{-10}. For both loss functions, almost no adjustments were necessary, especially for α close to 1; see Table 2.3. This suggests that although, the candidate for simultaneous optimality is not simultaneously optimal, it still is optimal with respect to several losses.

For the pair $(\mathbb{E}, \text{var})$, we chose weight functions $h_1(x) = 1/(x + 0.1)$ and $h_2(x) = \exp(-x/50 + 0.1)$. The reason for dividing by 50 was the scale of the weights to avoid numerical issues. The summand $+0.1$ was to avoid weights of zero. Again, the iteration was stopped when the loss given by (2.6) did not improve by more that 10^{-10}. Figure 2.7 displays the corresponding solutions obtained for a specific data set. The average number of iterations is displayed in Table 2.4.

Table 2.3: This table contains the average number of iterations for the two weight functions h_1, h_2 considered for the pair $(q_\alpha, \mathrm{ES}_\alpha)$.

			$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
$n = 10$	$\sigma = 3$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	h_1	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 100$	$\sigma = 3$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	h_1	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	h_1	0.03	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	h_1	0.03	0.04	0.04	0.01	0.00	0.00	0.00	0.00	0.00
		h_2	0.04	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$n = 500$	$\sigma = 3$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	h_1	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	h_1	0.03	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	h_1	0.04	0.05	0.06	0.01	0.00	0.00	0.00	0.00	0.00
		h_2	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 1000$	$\sigma = 3$	h_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	h_1	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	h_1	0.05	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	h_1	0.06	0.05	0.11	0.00	0.00	0.00	0.00	0.00	0.00
		h_2	0.05	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.4: The average number of iterations are displayed for the two weight functions h_1, h_2 considered for the pair $(\mathbb{E}, \text{var})$.

		$c = 0.5$	$c = 1$	$c = 3$	$c = 6$
$n = 10$	h_1	0.07	0.24	1.15	2.77
	h_2	0.01	0.06	0.94	2.79
$n = 100$	h_1	10.52	12.48	13.95	14.45
	h_2	2.90	6.52	13.12	12.73
$n = 500$	h_1	10.54	11.76	13.77	13.58
	h_2	6.68	11.05	14.40	8.78
$n = 1000$	h_1	9.16	10.05	11.37	12.58
	h_2	7.91	12.04	12.19	3.32

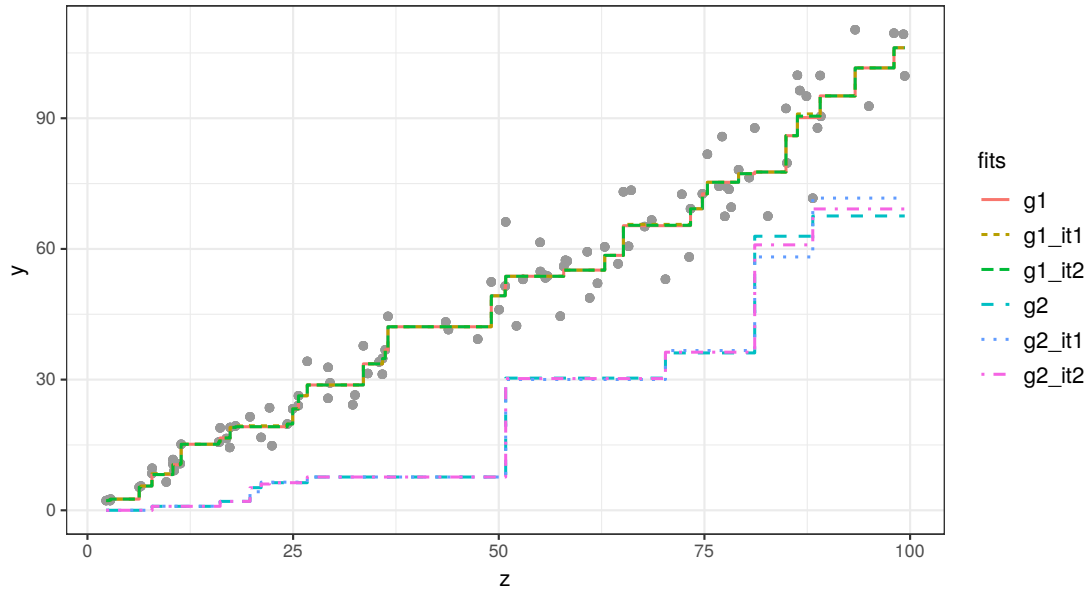

Figure 2.7: For a specific sample of size 100 the original fits (g1 and g2) are displayed in red and light blue respectively. The light green and dark blue fits (g1_it1 and g2_it1) correspond to the iterated versions of g_1 and g_2, respectively, with respect to weight function h_1. Finally, the dark green and the pink fits (g1_it2 and g2_it2) correspond to the iterated versions of g_1 and g_2, respectively, with respect to h_2. For h_1 the number of iterations was 13 and for h_2 a total of 5 iterations were necessary.

2.A Generalizations to partial orders

The results in this article can be generalized to partially ordered covariate sets. Let distribution P be the distribution of the random vector $(Z, Y) \in \mathcal{Z} \times \mathbb{R}$, where $\mathcal{Z}$ is a finite partially ordered set. We denote the partial order by $\preceq$. We aim now to minimize

the criterion

$$\int \tilde{L}(g_1(z), g_2(z), y)\, P(\mathrm{d}z, \mathrm{d}y) \\ = \int \Big(H(g_2(z)) + h(g_2(z)) \Big(L(g_1(z), y) - g_2(z) \Big) \Big) \, P(\mathrm{d}z, \mathrm{d}y) \quad (2.13)$$

among all increasing functions $g_1 \colon \mathcal{Z} \to \mathbb{R}$ and decreasing $g_2 \colon \mathcal{Z} \to \mathbb{R}$, that is, for $z \preceq z'$ we have $g_1(z) \leq g_1(z')$ and $g_2(z) \geq g_2(z')$. We call any minimizing pair a solution to the isotonic regression problem. Following Jordan et al. (2019), in order to accommodate the partially ordered set $\mathcal{Z}$, we introduce upper sets $x \subseteq \mathcal{Z}$ to replace single indices $i \in \{1, \ldots, n+1\}$. Set x is said to be an *upper set* if $z \in x$ and $z \preceq z'$ implies $z' \in x$. We denote $P_x(A) = P((x \times \mathbb{R}) \cap A)$ for any $A \in \mathcal{P}(\mathcal{Z}) \otimes \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. Let $\mathcal{X}$ consist of all admissible superlevel sets for an increasing function g imposed by the partial order on $\mathcal{Z}$.

As in the case of total orders, keeping either g_1 or g_2 fixed, we can find the optimal solution to (2.13) with respect to the other component.

Proposition 2.A.1. *(a) Let $g_1 : \mathcal{Z} \to \mathbb{R}$ be given. Then, the optimal antitonic solution $\hat{g}_2$ of (2.6) corresponding to $\hat{g}_1$ is given by*

$$\hat{g}_2(z) = - \min_{x' : z \notin x'} \max_{x \supsetneq x'} -\mathbb{E}(\bar{P}_{x \setminus x'}) = - \max_{x : z \in x} \min_{x' \subsetneq x} -\mathbb{E}(\bar{P}_{x \setminus x'}),$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(g_1(z_i), y_i), \ldots, L(g_1(z_j), y_j)$.

(b) Let $g_2 : \mathcal{Z} \to \mathbb{R}$ be given. Then, any optimal isotonic solution $\hat{g}_1$ of (2.6) with g_2 fixed satisfies

$$\min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-(P^w_{x \setminus x'}) \leq \hat{g}_1(z) \leq \max_{x : z \in x} \min_{x' \subsetneq x} T^+(P^w_{x \setminus x'}),$$

where $P^w_{x \setminus x'}$ is the weighted empirical distribution of y with $z \in x \setminus x'$ and weights proportional to $h(g_2(z)), z \in \mathcal{Z}$.

Proof. Follows with the same argument as for total orders. □

As in Section 2.3.2, we need to introduce some notation for the investigations ahead. In the following, we denote an optimal solution on the subset $\bar{x} \subseteq \mathcal{Z}$ by $\hat{g}_{1;\bar{x}}$ and by $\hat{g}_1|_{\bar{x}}$ we denote the optimal solution on the original set redistricted to $\bar{x}$.

Thinking in terms of superlevel sets, Lemma 2.3.3 states that $\hat{g}_1|_{\mathcal{Z} \setminus \bar{x}} \leq \hat{g}_{1;\mathcal{Z} \setminus \bar{x}}$ for any $\bar{x} \in \mathcal{X}$. Equivalently, $\hat{g}_1|_{\bar{x}} \geq \hat{g}_{1;\bar{x}}$.

Lemma 2.A.2. *Let $\bar{x} \in \mathcal{X}$ and assume that*

$$\hat{g}_1(z) := \min_{x': z \notin x'} \max_{x \supsetneq x'} T^\lambda(P^w_{x \setminus x'}) = \max_{x: z \in x} \min_{x' \subsetneq x} T^\lambda(P^w_{x \setminus x'})$$

for some $\lambda \in [0,1]$. Then we have $\hat{g}_1|_{\bar{x}} \geq \hat{g}_{1;\bar{x}}$.

Proof. It suffices to notice that

$$\hat{g}_1|_{\bar{x}}(z) = \max_{x: z \in x} \min_{x' \subsetneq x} T^\lambda(P^w_{x \setminus x'}) \geq \max_{\substack{x \in \mathcal{X};\, x \subseteq \bar{x};\\ z \in x}} \min_{x' \in \mathcal{X};\, x' \subsetneq x} T^\lambda(P^w_{x \setminus x'}) = \hat{g}_{1;\bar{x}}(z). \qquad \square$$

Let us recall the following observations made in Jordan et al. (2019). For fixed weights w, we can minimize

$$\int_{x \times \mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y), \quad \text{for all } \eta \in \mathbb{R} \tag{2.14}$$

among all admissible superlevel sets x for an increasing function $g_1 : \mathcal{Z} \to \mathbb{R}$ to obtain an optimal solution to (2.13). The search for the optimal superlevel set x needs to be conducted for every $\eta \in \mathbb{R}$. Again there is a one-to-one correspondence between admissible superlevel sets and optimal solutions. Instead of an increasing function $\iota : \mathbb{R} \to \{1, \ldots, n+1\}$ with $\iota(\eta) \in I_{1:n}(\eta)$ for all η, we now have a decreasing function $\xi : \mathbb{R} \to \mathcal{Z}$, in the sense that $\xi(\eta') \subseteq \xi(\eta)$ for $\eta' > \eta$. Moreover, it should hold that $\xi(\eta) \in X_{\mathcal{Z}}(\eta)$ for all $\eta \in \mathbb{R}$, where $X_{\mathcal{Z}}(\eta) \subseteq \mathcal{X}$ denotes the set of all superlevel sets minimizing (2.14). Then the correspondence between an optimal solution $\hat{g}_1$ and $\xi(\eta)$ is given by

$$\inf\{\eta : z \notin \xi(\eta)\} = \hat{g}_1(z) = \max\{\eta : z \in \xi(\eta)\}.$$

The next result is the generalization of Lemma 2.3.4 to partial orders.

Lemma 2.A.3. *Let $\bar{x} \in \mathcal{X}$. We have that $X_{\mathcal{Z}}(\eta) \cap (\mathcal{Z} \setminus \bar{x}) \subseteq X_{\mathcal{Z} \setminus \bar{x}}(\eta)$, where $X_{\bar{x}}(\eta)$ is the set of minimizing superlevel sets for the isotonic regression problem (2.13) on the subsample $(z, y), z \in \bar{x} \subseteq \mathcal{Z}$.*

Proof. Let $x' \in X_{\mathcal{Z}}(\eta) \cap (\mathcal{Z} \setminus \bar{x})$ for some $\eta \in \mathbb{R}$. Therefore, the function

$$t_\eta : \mathcal{Z} \to \mathbb{R}, x \mapsto \int_{x \times \mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y)$$

has a minimum at x'. We can write

$$\int_{x \times \mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) = \int_{x \cap (\mathcal{Z} \setminus \bar{x}) \times \mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) + \int_{x \cap \bar{x} \times \mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y).$$

Hence, $t_\eta|_{\mathcal{Z}\setminus\bar{x}}$ has a minimum at x' and thus $x' \subseteq X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. If t_η has a minimum in $x = \bar{x}$, then

$$t_\eta(x) - \int_{x\cap\bar{x}\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) \geq 0,$$

with equality in $x = \bar{x}$. Thus, $\emptyset \in X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. □

Let us generalize Proposition 2.3.7 to partial orders.

Proposition 2.A.4. *For fixed g_2, corresponding $\hat{g}_1^-$ and any increasing $\hat{g}_1$ we have*

$$\int \tilde{L}(g_1^-(z), g_2(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \leq \int \tilde{L}(g_1(z), g_2(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y).$$

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}^-$ denote the partition of $\mathcal{Z}$ corresponding to $\hat{g}_1$ and $\hat{g}_1^-$, respectively. By Lemma 2.2.1, it suffices to show that for all $\eta \in \mathbb{R}$

$$\int S_{\eta,2}(\hat{g}_1^-(z), g_2(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \leq \int S_{\eta,2}(\hat{g}_1(z), g_2(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y).$$

For the latter, it suffices to show that for all $\bar{x} \in \mathcal{X}$

$$\int_{\bar{x}\times\mathbb{R}} L(\hat{g}_1^-(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \leq \int_{\bar{x}\times\mathbb{R}} L(\hat{g}_1(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y).$$

Again it suffices to consider $\hat{g}_1$ with superlevel stets in $\cup_\eta X(\eta)$ and again we will prove the converse. In other words, for all $\bar{x} \in \mathcal{X}$ we have

$$\int_{\mathcal{Z}\setminus\bar{x}\times\mathbb{R}} L(\hat{g}_1(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \leq \int_{\mathcal{Z}\setminus\bar{x}\times\mathbb{R}} L(\hat{g}_1^-(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \tag{2.15}$$

If $\mathcal{Z} \setminus \bar{x} = Q_1 \cup \cdots \cup Q_i$, $Q_1, \ldots, Q_i \in \mathcal{Q}$, Lemma 2.A.3 implies that $\hat{g}_1|_{\mathcal{Z}\setminus\bar{x}}$ is optimal on $(z, y), z \in \mathcal{Z} \setminus \bar{x}$. Thus, (2.15) holds trivially. If there exists no sequence of partition elements such that $\mathcal{Z} \setminus \bar{x} = Q_1 \cup \cdots \cup Q_i$ we distinguish two cases.

Case 1: If $\mathcal{Z} \setminus \bar{x} = Q_1^- \cup \cdots \cup Q_{i^-}^-$, $Q_1^-, \ldots, Q_{i^-}^- \in \mathcal{Q}^-$ Lemma 2.A.2 implies that

$$\hat{g}_1^-|_{\mathcal{Z}\setminus\bar{x}} = \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^- \leq \hat{g}_1|_{\mathcal{Z}\setminus\bar{x}} \leq \hat{g}_1^+|_{\mathcal{Z}\setminus\bar{x}} \leq \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^+$$

Moreover, by Lemma 2.A.3, $X_{\mathcal{Z}}(\eta) \cap (\mathcal{Z} \setminus \bar{x}) \subseteq X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. Hence $\xi|_{\mathcal{Z}\setminus\bar{x}}(\eta) \in X_{\mathcal{Z}\setminus\bar{x}}(\eta)$ for all $\eta \in \mathbb{R}$, where $\xi : \mathbb{R} \to \mathcal{Z}$ is the function imposing the score-minimizing superlevel sets corresponding to $\hat{g}_1$. Thus, by Proposition 4.5 in Jordan et al. (2019) $\hat{g}_1|_{\mathcal{Z}\setminus\bar{x}}$ is an optimal solution to the isotonic regression problem on $(z, y), z \in \mathcal{Z} \setminus \bar{x}$.

Case 2: It remains to consider the case where no sequence of partition elements such that $\mathcal{Z} \setminus \bar{x} = Q_1^- \cup \cdots \cup Q_{i_-}^-$ exists. Note that $\hat{g}_1$ is optimal for all $z \in \mathcal{Z} \setminus \bar{x}$ with

$\hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z) \leq \hat{g}_1(z)$. Indeed, for those z, we have $\bar{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z) \leq \hat{g}_1(z) \leq \hat{g}^+_{1;\mathcal{Z}\setminus\bar{x}}(z)$, and can argue as in case 1. For $z \in \mathcal{Z}\setminus\bar{x}$ with $\hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z) > \hat{g}_1(z)$, we can argue similarly as in the proof of Proposition 2.3.7. For every $z \in \{z' \in \mathcal{Z}\setminus\bar{x} : \hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z') > \hat{g}_1(z')\}$ we have $z \in Q_{i+r}$, $r \in \{1,\dots,k\}$. Moreover, $\hat{g}_1$ is constant on every each Q_{i+r}, $r \in \{1,\dots,k\}$. With the same reasoning as in the proof of Proposition 2.3.7, we obtain that

$$\begin{aligned}\int_{Q^>_{i+r}\times\mathbb{R}} L(\hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) &\leq \int_{Q^>_{i+r}\times\mathbb{R}} L(c_i, y)\, P(\,\mathrm{d}z,\,\mathrm{d}y)\\ &\leq \int_{Q^>_{i+r}\times\mathbb{R}} L(c_i^-, y)\, P(\,\mathrm{d}z,\,\mathrm{d}y)\end{aligned}$$

for all $r \in \{1,\dots,k\}$, where $Q^>_{i+r} := Q_{i+r} \cap \{z \in \mathcal{Z}\setminus\bar{x} : \hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z) > \hat{g}_1(z)\}$. This implies the statement. □

Proposition 2.3.8 also translates directly to partial orders.

Proposition 2.A.5. *Assume that there exist $\hat{g}_1, \hat{g}_2 \colon \mathcal{Z} \to \mathbb{R}$ minimizing (2.13), then $\hat{g}_1^-(\cdot;\hat{g}_2)$, and the corresponding $\hat{g}_2^-(\cdot;\hat{g}_1^-(\cdot;\hat{g}_2))$ are also minimizers.*

Proof. The argument is the same as in the proof of Proposition 2.3.8. □

As in the case of total orders a simultaneously optimal solution may not necessarily exists, since $\hat{g}_2^-$ imposes additional constraints. Nonetheless, we are able to formulate a criterion so that simultaneous optimality is reached whenever the criterion is fulfilled. Let

$$\begin{aligned}\hat{g}_1(z) &= \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-(P_{x\setminus x'}) = \max_{x : z \in x} \min_{x' \subsetneq x} T^-(P_{x\setminus x'}),\\ \hat{g}_2(z) &= -\min_{x' : z \notin x'} \max_{x \supsetneq x'} -\mathbb{E}(\bar{P}_{x\setminus x'}) = -\max_{x : z \in x} \min_{x' \subsetneq x} -\mathbb{E}(\bar{P}_{x\setminus x'}),\end{aligned}$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(g_1(z), y)$, $z \in \mathcal{Z}$.

Proposition 2.A.6. *Let $\hat{g}_1^-$, $\hat{g}_2^-$ as defined above. A simultaneously optimal solution exists if and only if $\hat{g}_1^- = \hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}$ for all superlevel sets $\mathcal{Z}\setminus\bar{x}$, $\bar{x} \in \mathcal{X}$ assumed by $\hat{g}_2^-$.*

The reasoning behind this Proposition is analogous to the reasoning behind Proposition 2.3.9.

3 Forecasting value-at-risk and expected shortfall using isotonic regression

This chapter supplements our work done in Chapter 2 where we studied isotonic regression for functionals of elicitation complexity greater than one. In particular, we considered the pair value-at-risk (VaR) and expected shortfall (ES). This manuscripts provides insight into the performance of the use of isotonic regression to predict VaR and ES compared to other methods. Using a real world data set, we aim to compare the pool-adjacent-violators (PAVA) approach to state-of-the-art methods for risk prediction.

3.1 Setup and forecasting methods

In the spirit of Nolde and Ziegel (2017) we try to model the *negated percentage log-returns* of the NASDAQ composite index, that is, if v_t is the value of the index, measured in points, at time t and v_{t-1} is the value at time $t-1$ then the negated percentage log-return is defined by

$$-\log\left(\frac{v_{t-1}}{v_t}\right)\cdot 100.$$

It is assumed that the risk increases with higher volatility. Therefore, we consider the volatility index VXN of the NASDAQ composite index as an explanatory variable for the isotonic regression. More rigorously, we consider the pairs $(z_0, y_1), \ldots, (z_{n-1}, y_n)$, where $\{y_t\}_{t=1}^n$ is the time series of negated percentage log-returns and $\{z_t\}_{t=0}^{n-1}$ is the time series of the previous-day volatility. In particular, we assume that there is an isotonic relationship between the two time series. Since the sign convention for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ used by Nolde and Ziegel (2017) differs from the sign convention we use in Chapter 2, we adopt their sign convention for this chapter. This means that we look for $\hat{g}_1$, $\hat{g}_2 : \{z_0, \ldots, z_{n-1}\} \to \mathbb{R}$ isotonic. The simulation study in Section 2.4 suggests that for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, the candidate for simultaneous optimality is already optimal for some specific loss functions. We checked and found that this continues to hold for all fits and losses considered in this numerical study. Thus, for simplicity, we do not distinguish between the estimates obtained from the different losses considered for the evaluation as

they coincide. In particular, they coincide with the candidate for simultaneous optimality. We therefore estimate $\hat{g}_1$, $\hat{g}_2$ via

$$\hat{g}_1(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} T^-(P_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P_{i:j})$$
$$\hat{g}_2(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} \mathbb{E}(\bar{P}_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} \mathbb{E}(\bar{P}_{i:j}), \quad \ell = 1, \ldots, n,$$

where $P_{i:j}$ is the empirical distribution of $y_i, \ldots, y_j$, and $\bar{P}_{i:j}$ is the empirical distribution of $L(\hat{g}_1(z_{i-1}), y_i), \ldots, L(\hat{g}_1(z_{j-1}), y_j)$ where

$$L(x_1, y) = -\frac{1}{\alpha} \mathbb{1}\{y > x_1\}(x_1 - y) + x_1$$

due to the new sign convention. We denote this method by PAVA-vol since it can be efficiently calculated via the pool-adjacent-violators algorithm. For a detailed introduction; see Barlow et al. (1972).

The data set we consider for this numerical study is publicly available and has been downloaded from http://finance.yahoo.com. Our data set spans from January 2, 2003, to February 20, 2020. We split the data into a training and a test period. To increase insight we consider two different training periods separately. For both training sets, the test period spans from January 2, 2009, to February 20, 2020. This yields an out-of-sample size $N = 2802$. The shorter training period starts on January 2, 2003, ends on December 29, 2006 and has thus a sample size of $n = 765$. The longer training period also starts on January 2, 2003, but ends on December 31, 2008. Therefore, the training sample size in this case is $n = 1151$. Figure 3.1 displays the two different scenarios.

Naturally, in order to assess the performance of the isotonic regression approach, we would like to compare our method to some other approaches. For this purpose we have chosen to consider *filtered historical simulation* (FHS) and a rolling window approach.

Filtered historical simulation. Filtered historical simulation was also considered in Nolde and Ziegel (2017). Moreover, they kindly published their implementations online. We used and adapted their code to fit our setup. Like Nolde and Ziegel (2017) we assume that the time series of negated log-returns $\{y_t\}_{t \in \mathbb{N}}$ can be modeled as $y_t = \mu_t + \sigma_t x_t$, where $\{x_t\}_{t \in \mathbb{N}}$ is a sequence of independent and identically distributed random variables with zero mean and unit variance. Furthermore, μ_t, σ_t are assumed to be measurable with respect to the sigma-algebra $\mathcal{F}_{t-1}$ which is assumed to contain the information of the process $\{y_t\}_{t \in \mathbb{N}}$ that is available at time $t - 1$. In order to capture the dynamics necessary for financial time series, we fit an AR(1)-GARCH(1, 1) model to the time series of the negated percentage log-returns on the the training data. Moreover, we first used normally distributed innovations $\{x_t\}_{t \in \mathbb{N}}$ (denoted by n-FHS) and we secondly used skewed t-distributed (denoted by st-FHS) innovations $\{x_t\}_{t \in \mathbb{N}}$.

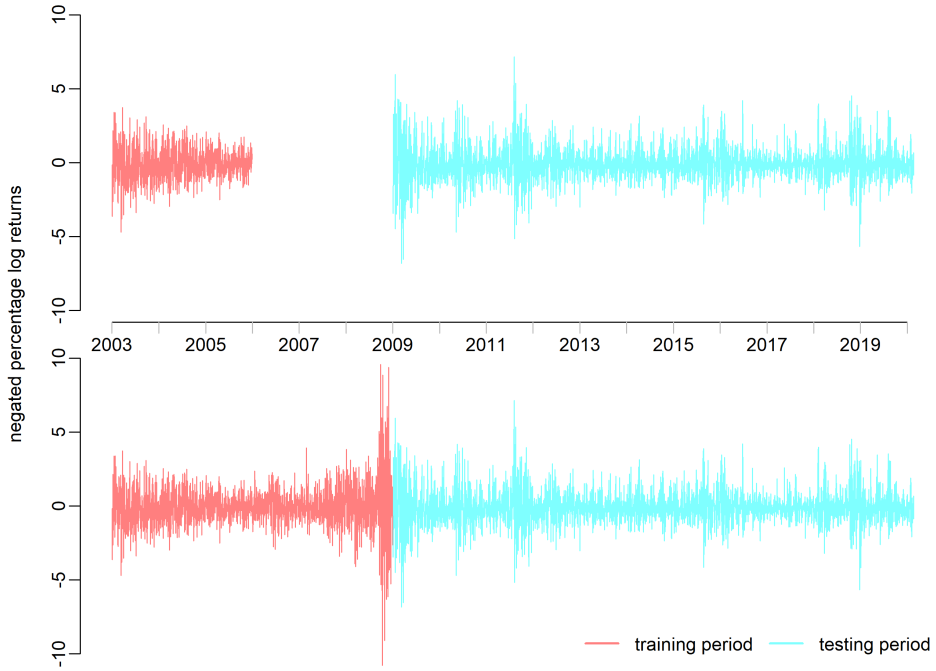

Figure 3.1: On the top the training data for the short training period (red) and the test data (blue) are drawn. Below the same is displayed for the long training period.

In the estimation process, one exploits that

$$\rho(\mathcal{L}(y_t|\mathcal{F}_{t-1})) = \hat{\mu}_t + \hat{\sigma}_t \hat{\rho}(\mathcal{L}(X))$$

where $\rho =$ VaR or ES and the random variable X has the same distribution as the innovations. In a first step, μ_t and σ_t are estimated via maximum likelihood under specific assumption on the distribution of the innovations x_t. Then, $\rho(\mathcal{L}(X))$ is estimated is estimated based on the sample of standardized residuals $\hat{x}_t = (y_t - \hat{\mu}_t)/\hat{\sigma}_t$. This is done by drawing a sample $\{\hat{x}_i^* : 1 \leq i \leq M\}$ of a large sample size M from the residuals $\{\hat{x}_t : 1 \leq t \leq n\}$. The value-at-risk is then the empirical α-quantile. The expected shortfall is the empirical version of the conditional expectation given that the residual exceeds the value-at-risk. Finally, using the above equality, one can estimate $\rho(\mathcal{L}(y_t|\mathcal{F}_{t-1}))$ by $\hat{\mu}_t + \hat{\sigma}_t \hat{\rho}(\mathcal{L}(Z))$ for $t = n+1, \ldots n+N$.

Rolling window. The second approach we consider for comparison is a rolling window approach as described in Patton et al. (2019). Hereby, the value-at-risk is estimated by the sample quantile on $\{y_s\}_{s=t-w}^{t-1}$ and the expected shortfall is estimated by the corresponding empirical expected shortfall. We chose a window of size $w = 250$. It should be noted, that the comparison with the rolling window approach is not completely fair. Both, in filtered historical simulation estimate and in isotonic regression, we do not have access to the observations y_t in the test sample during the estimation process.

Nonetheless, it can be revealing to compare the estimates with the rolling window approach. We denote this method by RW-250. Clearly, it is possible to apply isotonic regression and FHS in rolling window mode. However, due to the nature of the isotonic regression approach, the window size needs to be substantially lager than 250 to obtain a reliable estimate for the isotonic relationship between the volatility and the negated percentage log-return. For a fair comparison with FHS the same window size should be applied, leading to a higher computational cost. Furthermore, isotonic regression applied in rolling window mode does not lead to substantially better results since adding an additional data point does generally not yield a considerable change in the fit because the quantile fit is rather robust. Thus, we restricted ourselves to a full out-of-sample approach in order to gain the greatest understanding of the performance of the isotonic regression.

Now that all the methods have been briefly introduced, it remains to be mentioned that we will look at the three levels $\alpha \in \{0.9, 0.95, 0.99\}$.

3.1.1 Evaluation

The different forecasting methods are compared using the same measures as Nolde and Ziegel (2017). That is, for the VaR_α, the percentage of violations (% Viol.), the 1-homogeneous loss function

$$S_A(x_1, y) = \Big(1 - \alpha - \mathbb{1}\{y > x_1\}\Big)x_1 + \mathbb{1}\{y > x_1\}y$$

and the loss

$$S_B(x_1, y) = \Big(1 - \alpha - \mathbb{1}\{y > x_1\}\Big)\log x_1 + \mathbb{1}\{y > x_1\}\log y, \quad x_1 > 0,$$

which has 0-homogeneous loss differences. For the pair $(\text{VaR}_\alpha, \text{ES}_\alpha)$, we use

$$S_C\left(x_1, x_2, y\right) = \mathbb{1}\left\{y > x_1\right\}\frac{y - x_1}{2\sqrt{x_2}} + (1 - \alpha)\frac{x_1 + x_2}{2\sqrt{x_2}}$$

and the alternative choice

$$S_D\left(x_1, x_2, y\right) = \mathbb{1}\left\{y > x_1\right\}\frac{y - x_1}{x_2} + (1 - \alpha)\left(\frac{x_1}{x_2} - 1 + \log\left(x_2\right)\right),$$

which has 0-homogeneous loss differences.

Table 3.1: The above table contains the results of the in-sample evaluation for the short training period. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA	1.315	**10.847**	**0.183**	**0.056**	1.374	**0.136**	**0.063**
	n-FHS	1.827	7.275	0.206	0.065	2.308	0.142	0.067
	st-FHS	1.677	8.862	0.203	0.064	2.238	0.141	0.066
$\alpha = 0.95$	PAVA	1.754	5.423	**0.104**	**0.035**	1.772	**0.072**	**0.037**
	n-FHS	2.116	3.439	0.117	0.039	2.551	0.075	0.039
	st-FHS	2.032	**4.894**	0.116	0.039	2.526	0.075	0.039
$\alpha = 0.99$	PAVA	2.272	0.661	**0.023**	**0.008**	2.273	**0.015**	**0.008**
	n-FHS	2.746	**0.794**	0.028	0.010	3.010	0.017	0.010
	st-FHS	3.043	0.132	0.030	0.010	3.135	0.017	0.010

Table 3.2: The above table contains the results of the in-sample evaluation for the long training period. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA	1.394	13.964	**0.228**	0.073	1.493	0.152	0.088
	n-FHS	1.661	10.523	**0.228**	**0.070**	2.301	**0.147**	**0.072**
	st-FHS	1.691	**10.457**	0.230	0.071	2.302	**0.147**	0.073
$\alpha = 0.95$	PAVA	1.818	8.471	0.138	0.049	1.869	0.084	0.056
	n-FHS	2.027	5.559	**0.132**	**0.043**	2.672	**0.079**	**0.043**
	st-FHS	2.152	**4.964**	0.133	**0.043**	2.745	**0.079**	0.044
$\alpha = 0.99$	PAVA	2.932	2.581	0.040	0.015	2.943	0.021	0.016
	n-FHS	3.143	0.397	**0.034**	**0.011**	4.212	**0.018**	**0.011**
	st-FHS	2.995	**0.596**	**0.034**	**0.011**	3.313	**0.018**	**0.011**

3.2 Results

3.2.1 In-sample performance

First of all, we evaluated the two-stage isotonic regression approach and the filtered historical simulation in-sample. Tables 3.1 and 3.2 contain the results for the short and long training period, respectively.

What stands out directly is that the isotonic regression approach leads to substantially more violations (% Viol.) than desired for the long training period. When focusing on the training set, it is apparent from Figure 3.2 that while performing well for large volatility our method tends to underestimate the risk when the volatility is small. This could implicate that our method struggles with heavy-tail data.

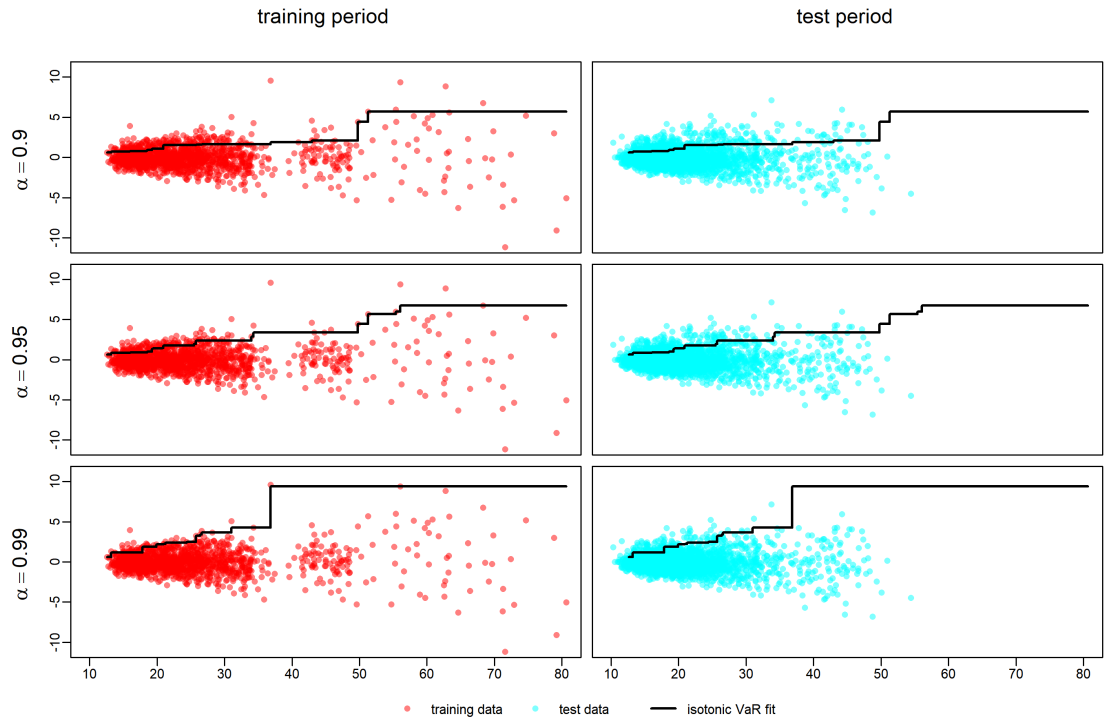

Figure 3.2: The above picture contains the data and fit based on the long training period. In red the training data for the two-stage isotonic regression approach is displayed. The blue points correspond to the test set. The black line is the isotonic regression fit based on the long training period. One can see that for similar values for the volatility the corresponding negated percentage log-returns are more extreme in the test set.

Going back to the Tables 3.1 and 3.2, we see that for the short training period our approach dominates n-FHS and st-FHS regarding the loss functions S_A-S_D. However, for the longer training period, our method becomes dominated by the filtered historical simulation. Theoretically our method should lead to an optimal isotonic in-sample fit for the scores S_A, S_B, S_C and S_D. Therefore, the lack in performance could hint that the isotonicity assumption is in fact violated.

3.2.2 Out-of-sample performance

Of course, the ultimate goal of this data example is the prediction of the value-at-risk and the expected shortfall. Thus, we turn our attention to the out-of-sample performance of the isotonic regression approach. Tables 3.3 and 3.4 display the results of the short and long training period, respectively.

Out-of-sample, the isotonic regression approach cannot keep up with the filtered historical simulation and the rolling-window approach. The PAVA approach gets nearly always

Table 3.3: The results of the out-of-sample evaluation for the short training period are displayed. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA-vol	1.109	12.027	**0.208**	**0.062**	1.160	0.148	0.086
	n-FHS	1.463	8.351	**0.208**	0.063	1.914	**0.143**	**0.069**
	st-FHS	1.532	7.637	0.210	0.064	2	**0.143**	**0.069**
	RW-250	1.454	**8.922**	0.221	0.071	2.225	0.148	0.077
$\alpha = 0.95$	PAVA	1.460	7.459	0.132	0.045	1.477	0.083	0.055
	n-FHS	1.781	5.389	**0.130**	**0.044**	2.192	**0.080**	**0.047**
	st-FHS	1.833	**5.175**	**0.130**	**0.044**	2.229	**0.080**	**0.047**
	RW-250	2.025	4.604	0.142	0.049	2.793	0.084	0.052
$\alpha = 0.99$	PAVA	1.900	4.461	0.051	0.019	1.901	0.025	0.023
	n-FHS	2.583	2.213	**0.039**	0.014	2.712	**0.020**	**0.014**
	st-FHS	2.684	1.999	**0.039**	**0.013**	2.777	**0.020**	**0.014**
	RW-250	3.609	**1.106**	0.045	0.015	3.328	0.021	0.015

Table 3.4: The results of the out-of-sample evaluation for the long training period are displayed. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA-vol	1.055	13.312	0.207	0.062	1.141	0.148	0.087
	n-FHS	1.416	8.637	**0.202**	**0.059**	2.049	**0.139**	**0.063**
	st-FHS	1.435	8.672	0.204	0.061	2.025	0.140	0.064
	RW-250	1.454	**8.922**	0.221	0.071	2.225	0.148	0.077
$\alpha = 0.95$	PAVA-vol	1.354	9.350	0.135	0.050	1.401	0.086	0.065
	n-FHS	1.782	5.460	**0.122**	**0.040**	2.365	**0.077**	**0.041**
	st-FHS	1.891	**4.640**	0.123	0.041	2.457	**0.077**	**0.041**
	RW-250	2.025	4.604	0.142	0.049	2.793	0.084	0.052
$\alpha = 0.99$	PAVA-vol	2.132	4.711	0.051	0.022	2.144	0.026	0.027
	n-FHS	2.742	0.892	**0.031**	**0.011**	3.320	**0.017**	**0.011**
	st-FHS	2.685	1.392	0.032	**0.011**	3.089	0.018	**0.011**
	RW-250	3.609	**1.106**	0.045	0.015	3.328	0.021	0.015

dominated by n-FHS, st-FHS and RW-250. As we have already seen when looking at the in-sample performance the percentage of violations of the PAVA approach is off. Additionally to the possible causes previously discussed this is probably amplified because the training data and the test data differ in terms of the scale of the data; see Figure 3.2. For similar volatility values the corresponding log-returns in the test set tend to be more extreme. Thus, the isotonic regression approach tends to underestimate the risk.

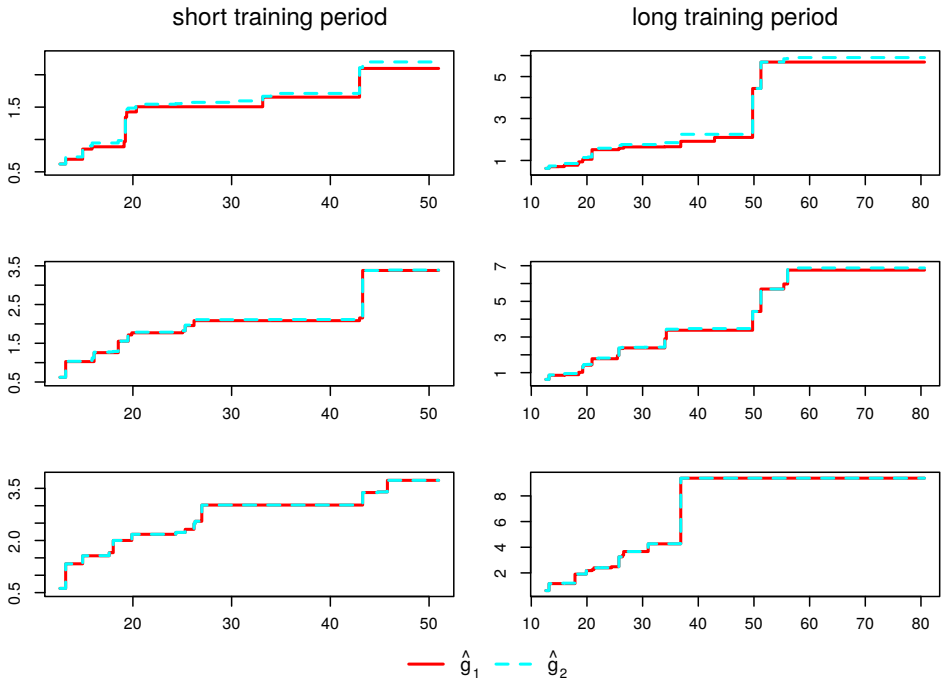

Figure 3.3: The isotonic fits $\hat{g}_1$ and $\hat{g}_2$ are displayed for all levels α and both training periods.

What also stands out is that the isotonic regression approach seems to have more issues to keep up with the ES predictions than with the VaR predictions of the other methods. Figure 3.3 contains the isotonic fits for all levels α and both training period. It is apparent that both fits are fairly similar. The reason for this lies in the transformation that is applied on the data before the isotonic regression for ES is performed. In other words, $\hat{g}_2$ is the optimal isotonic fit with respect to the mean on the data $(z_{i-1}, L(\hat{g}_1(z_{i-1}), y_i))$, $i \in \{1, \dots, n\}$, whereby $(z_0, y_1), \dots, (z_{n-1}, y_n)$ is the original data and $\hat{g}_1$ is the isotonic fit for VaR. Adapting to the sign convention used in Nolde and Ziegel (2017) and this chapter, the loss L is given by

$$L(x, y) = -\frac{1}{\alpha}\mathbb{1}\{x < y\}(x - y) + x.$$

But because $\hat{g}_1$ is the isotonic fit for the VaR (α-quantile) it rarely happens that $\mathbb{1}\{\hat{g}_1(z_{i-1}) < y_i\} \neq 0$ for values of α close to one. Hence, $(z_{i-1}, L(\hat{g}_1(z_{i-1}), y_i))$ is already nearly isotonic. Thus, only few modifications to the data are needed to obtain an isotonic fit $\hat{g}_2$. Therefore, $\hat{g}_2$ is close to $\hat{g}_1$ and so are the predictions.

3.3 Possible improvements

In this section we introduce and compare different possible improvements of the isotonic regression approach. All results are contained in Tables 3.5 and 3.6. These tables contain the out-of-sample comparisons.

Subbagging. As a first option, we apply subbagging to obtain new estimates $\hat{g}_1$ and $\hat{g}_2$. Hereby, we take a subsample S_1 of our training data of size $n/2$ and calculate the optimal fit for VaR on S_1, denoted by $\hat{g}_1|_{S_1}$. This process is repeated $M = 100$ times. Then, we let $\hat{g}_1 = M^{-1} \sum_{i=1}^{M} \hat{g}_1|_{S_i}$ and calculate the corresponding $\hat{g}_2$. We denote this modification by PAVA-vol-bag. In Tables 3.5 and 3.6, we see that applying subbagging does improve the performance, however not enough to truly compete with the reference methods.

Choose a different covariate. When we considered the in-sample performance, we noted that the isotonicity assumption may be violated. Thus, it could be reasonable to consider a different covariate instead. To this end, we choose the lag return, that is, the absolute value of the previous-day negated percentage log-return. We denote this method by PAVA-lag. The results show that this covariate yields better results in terms of %Viol over all levels α. Moreover, for $\alpha = 0.99$ it also yields better average scores than PAVA-vol. This is, however, not the case for $\alpha \in \{0.90, 0.95\}$. For $\alpha = 0.99$ the performance is comparable to the one of RW-250 but it can still not keep up with the filtered historical simulation. We also applied subbagging for PAVA-lag but there was no notable improvement, so that these results are not displayed.

Increase the training data. Because our isotonic regression approach relies on few assumptions, it is possible to add another data set to the training data, as long as it is on the same scale and relies on the same isotonic relationship. To increase our training sample size, we add the negated percentage log-returns from the S&P500 composite index together with its previous-day volatility index VIX as an explanatory variable to our data sets. We perform the original two-stage isotonic regression approach on the larger training set and also consider subbagging on this training set. The methods are denoted by PAVA-vol+ and PAVA-vol-bag+, respectively. For the long training period, PAVA-vol-bag+ performs better than the original PAVA-vol. Unfortunately, the improvement is not substantial enough to compete with the filtered historical simulation.

Add a second dependent variable. Henzi et al. (2019) introduced a nonparametric method called *isotonic distributional regression* (IDR) to estimate conditional distribution subject to isotonicity constraints on the covariate space. Isotonic quantile regression is a special case thereof. In contrast to the pool-adjacent-violators algorithm, we can consider

multiple explanatory variables when using IDR and still obtain a simultaneously optimal solution. This enables us to consider multiple covariates in the first stage of our isotonic regression approach. That is, to find the optimal $\hat{g}_1$ we can consider multiple covariates. To this end, we consider the volatility as well as the lag return as covariates and denote this method by IDR-vol-lag. Furthermore, we also apply this method to the training set enlarged with the S&P500 data and denote this variation as IDR-vol-lag+. For the short training period IDR-vol-lag and IDR-vol-lag+ tend to perform worse than the other methods considered. For the long training period, the results are mixed. Sometimes they do perform better than PAVA-vol-bag+ and sometimes not. However, they outperform PAVA-vol. This supports our suspicion that the relationship between the volatility and the negated percentage log-returns is not fully isotonic and thus more information helps in the estimation. However, the performance is still worse that for the methods we considered for reference.

3.4 Conclusions

We have compared estimation of the negated percentage log-returns via isotonic regression to some widely used approaches. The results lead to the conclusion that the two-stage isotonic regression cannot compete with the other methods in terms of performance. We have seen that this could be due to violation of the isotonicity assumption. Nonetheless, we are still surprised by the performance reached with so few underlying assumptions. One disadvantage of the isotonic regression approach is that it heavily relies on the training data and thus is not as flexible. But we are hopeful that there are important data sets where our method may be able to keep up with state-of-the art prediction methods for bivariate functionals.

Table 3.5: This table contains the results for the short training period for all methods discussed in this Chapter. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA-vol	1.109	12.027	0.208	0.062	1.160	0.148	0.086
	PAVA-vol-bag	1.109	12.455	0.206	0.061	1.170	0.146	0.084
	PAVA-lag	1.424	8.458	0.219	0.072	1.475	0.150	0.088
	PAVA-vol+	0.923	15.667	0.212	0.068	1.006	0.153	0.104
	PAVA-vol-bag+	1.047	13.205	0.207	0.062	1.109	0.148	0.088
	IDR-vol-lag	1.136	12.491	0.208	0.064	1.139	0.151	0.102
	IDR-vol-lag+	1.136	12.491	0.208	0.064	1.139	0.151	0.102
	n-FHS	1.463	8.351	0.208	0.063	1.914	0.143	0.069
	st-FHS	1.532	7.637	0.210	0.064	2	0.143	0.069
	RW-250	1.454	8.922	0.221	0.071	2.225	0.148	0.077
$\alpha = 0.95$	PAVA-vol	1.460	7.459	0.132	0.045	1.477	0.083	0.055
	PAVA-vol-bag	1.431	8.173	0.133	0.047	1.460	0.084	0.058
	PAVA-lag	1.743	6.174	0.139	0.049	1.767	0.085	0.057
	PAVA-vol+	1.304	9.672	0.136	0.049	1.335	0.086	0.063
	PAVA-vol-bag+	1.377	8.530	0.134	0.046	1.402	0.085	0.058
	IDR-vol-lag	1.411	8.423	0.138	0.050	1.411	0.088	0.070
	IDR-vol-lag+	1.411	8.423	0.138	0.050	1.411	0.088	0.070
	n-FHS	1.781	5.389	0.130	0.044	2.192	0.080	0.047
	st-FHS	1.833	5.175	0.130	0.044	2.229	0.080	0.047
	RW-250	2.025	4.604	0.142	0.049	2.793	0.084	0.052
$\alpha = 0.99$	PAVA-vol	1.900	4.461	0.051	0.019	1.901	0.025	0.023
	PAVA-vol-bag	1.794	5.567	0.055	0.024	1.805	0.028	0.030
	PAVA-lag	2.502	3.069	0.046	0.016	2.504	0.023	0.019
	PAVA-vol+	1.854	4.711	0.053	0.020	1.858	0.026	0.024
	PAVA-vol-bag+	1.857	4.889	0.053	0.020	1.861	0.026	0.023
	IDR-vol-lag	1.704	6.103	0.063	0.027	1.704	0.032	0.038
	IDR-vol-lag+	1.704	6.103	0.063	0.027	1.704	0.032	0.038
	n-FHS	2.583	2.213	0.039	0.014	2.712	0.020	0.014
	st-FHS	2.684	1.999	0.039	0.013	2.777	0.020	0.014
	RW-250	3.609	1.106	0.045	0.015	3.328	0.021	0.015

Table 3.6: The above table contains the results for the short training period for all methods discussed. The second column ($\overline{\text{VaR}}$) reports the average value-at-risk forecasts and the sixth column ($\overline{\text{ES}}$) reports the average expected shortfall forecasts.

		$\overline{\text{VaR}}$	% Viol.	$\overline{S_A}$	$\overline{S_B}$	$\overline{\text{ES}}$	$\overline{S_C}$	$\overline{S_D}$
$\alpha = 0.90$	PAVA-vol	1.055	13.312	0.207	0.062	1.141	0.148	0.087
	PAVA-vol-bag	1.097	12.991	0.205	0.062	1.188	0.146	0.085
	PAVA-lag	1.595	7.281	0.224	0.074	1.685	0.151	0.086
	PAVA-vol+	0.902	16.417	0.213	0.071	1.034	0.154	0.108
	PAVA-vol-bag+	1.069	13.348	0.205	0.062	1.173	0.147	0.086
	IDR-vol-lag	1.263	10.385	0.205	0.060	1.263	0.146	0.084
	IDR-vol-lag+	1.263	10.385	0.205	0.060	1.263	0.146	0.084
	n-FHS	1.416	8.637	0.202	0.059	2.049	0.139	0.063
	st-FHS	1.435	8.672	0.204	0.061	2.025	0.140	0.064
	RW-250	1.454	8.922	0.221	0.071	2.225	0.148	0.077
$\alpha = 0.95$	PAVA-vol	1.354	9.350	0.135	0.050	1.401	0.086	0.065
	PAVA-vol-bag	1.452	8.244	0.129	0.046	1.500	0.083	0.057
	PAVA-lag	2.049	4.675	0.143	0.050	2.099	0.086	0.056
	PAVA-vol+	1.259	11.420	0.141	0.056	1.332	0.090	0.075
	PAVA-vol-bag+	1.450	8.280	0.130	0.045	1.504	0.083	0.056
	IDR-vol-lag	1.585	6.888	0.132	0.046	1.585	0.084	0.059
	IDR-vol-lag+	1.585	6.888	0.132	0.046	1.585	0.084	0.059
	n-FHS	1.782	5.460	0.122	0.040	2.365	0.077	0.041
	st-FHS	1.891	4.640	0.123	0.041	2.457	0.077	0.041
	RW-250	2.025	4.604	0.142	0.049	2.793	0.084	0.052
$\alpha = 0.99$	PAVA-vol	2.132	4.711	0.051	0.022	2.144	0.026	0.027
	PAVA-vol-bag	2.134	4.390	0.046	0.022	2.154	0.025	0.026
	PAVA-lag	3.251	1.535	0.045	0.015	3.268	0.022	0.017
	PAVA-vol+	2.222	4.247	0.050	0.019	2.236	0.025	0.023
	PAVA-vol-bag+	2.341	3.176	0.042	0.015	2.351	0.022	0.017
	IDR-vol-lag	2.263	3.854	0.051	0.020	2.263	0.026	0.026
	IDR-vol-lag+	2.263	3.854	0.051	0.020	2.263	0.026	0.026
	n-FHS	2.742	0.892	0.031	0.011	3.320	0.017	0.011
	st-FHS	2.685	1.392	0.032	0.011	3.089	0.018	0.011
	RW-250	3.609	1.106	0.045	0.015	3.328	0.021	0.015

4 Pareto-optimal parameters in linear regression problems

Anja Mühlemann and Johanna F. Ziegel

Abstract. Parametric regression models tend to create a false sense of security. In fact, parametric models are often misspecified in that they are a simplification of the reality. This misspecification may not necessarily impact our interpretations and predictions but it can. It is therefore not surprising that the development of methods to detect misspecification as well as methods less sensitive to misspecification has gained a lot of attention recently. We introduce the concept of Pareto-optimal parameters, that is model parameters that are not strictly dominated by any other model parameter within a prespecified class of loss functions. Pareto-optimal parameters coincide with the true model parameters under correct specification. We show how the set of Pareto-optimal parameters can be used to detect misspecification and how it can be interpreted as a measure of model uncertainty. We discuss the calculation of the Pareto-optimal set on the population as well as on the sample level in case of isotonic regression. Finally, we put our method to the test by considering two data sets.

Acknowledgments. We would like to thank our colleagues at the University of Bern for many valuable discussions. Moreover, we gratefully acknowledge financial support from the Swiss National Science Foundation.

4.1 Introduction

Ideally, a forecasting model captures reality in its entirety. However, in most applications models are simplifications and can thus only capture a part of reality. This model misspecification impacts, although not always, our estimates and predictions. It is therefore essential to develop statistical methods to detect misspecification. Nevertheless, since a correctly specified model may not be available, there is a need for methods that perform reasonably well even in the presence of misspecification.

In econometrics, forecasters find themselves in a situation where they have to issue a forecast on the future value of a variable. The value of this variable however depends on future choices of market participants. Therefore, their models are likely to be misspecified by the time the true value is assumed. Feiler and Ajdler (2019) suggest to incorporate relations among competing models in the estimation process to reduce uncertainty.

In parameter estimation, uncertainty is often addressed by calculating a confidence interval instead of a point estimate. Hansen et al. (2011) adopt this idea and introduce a model confidence set that is constructed such that it contains the best model within a class of models with a given confidence.

Grünwald and Roos (2019) considers an Occam's razor point of view in suggesting that for a given data set the best explanation is provided by the shortest description of the data. This approach optimizes the trade-off between goodness-of-fit and model complexity.

Another approach would be to develop estimation methods requiring less model assumptions so that the chance of misspecification is minimized. For mean estimation, Holland (2019) introduces a class of mean estimators with finite variance being the only assumption.

Standard Bayesian inference is known to be susceptible to model misspecification. It is therefore not surprising that developing methods less sensitive to misspecification has gained a lot of attention recently. Huggins and Miller (2019) show that using bootstrap to obtain bagged posteriors (BayesBag) has better predictive accuracy than the standard Bayesian approach when misspecification is present. And even in a correctly specified scenario BayesBag produced better or equally good results. Huggins and Miller (2019) also introduce a so-called *mismatch index* allowing for model criticism. Thomas and Corander (2019) take a different approach. Instead of using a bagged posterior when misspecification is suspected they perform a Bayes' update with the likelihood raised to a power $t \in [0, 1]$. This approach is known as tempering and helps avoid convergence to a poor model in the presence of misspecification. To recognize the presence of misspecification and to choose a suitable tempering parameter t, they suggest to train a probabilistic classifier to discriminate between the observed data and some simulated data. Another approach named *focused Bayesian prediction* is proposed by Loaiza-Maya

et al. (2019). They replace the standard Bayesian up-date by a criterion that captures a user-specified measure of predictive accuracy.

However, Bayesian inference is by far not the only approach sensitive to model misspecification. For linear models, Buja et al. (2019) intuitively explain how in the presence of non-linearity the covariates can no longer be treated as fixed. Randomness of the covariates, however, affects the parameter estimates and creates sampling variability.

When a parametric model is desired, the general approach is to minimize the loss between the model and the observations to obtain the optimal parameter. In ordinary least squares (OLS) regression, the aim is to estimate the conditional mean which is done by minimizing the squared error.

The reason to minimize the squared error is that, under correct specification on the population level, the squared error is a *consistent* loss function for the mean in the sense of Gneiting (2011). Let $I \subseteq \mathbb{R}$ be an interval and let $\mathcal{P}$ be a class of probability distributions on I. Then a loss function L is said to be consistent for the mean if $\mathbb{E}L(\mathbb{E}(Y), Y) \leq \mathbb{E}L(x, Y)$ for all $x \in I, P \in \mathcal{P}$, where Y is a random variable with distribution P.

Savage (1971) showed that, under mild regularity conditions, an entire class $\mathcal{L}$ of loss functions, the Bregman class, is consistent for the mean. But consistency of the Bregman loss functions for the mean only ensures that the true parameter of a correctly specified model minimizes the Bregman loss the population level. In the case of a misspecified model the optimal parameters vary substantially depending on the Bregman loss (Patton, 2020).

This is where we would like to tie in. Adapting the definition of forecast dominance in Ehm et al. (2016), we say that parameter θ_1 is *dominated* by parameter θ_2 if

$$\mathbb{E}L_\phi(m(X;\theta_2), Y) \leq \mathbb{E}L_\phi(m(X;\theta_1), Y)$$

for all $L_\phi \in \mathcal{L}$, where $m(x;\theta)$ denotes a parametric model for the conditional mean g with model parameter θ. It is *strictly dominated* if the above inequality furthermore is strict for some $L_\phi \in \mathcal{L}$. We will show that if one parameter θ dominates all other parameters, then indeed $g(\cdot) = m(\cdot\,;\theta)$. However, in general there is no parameter dominating all other parameters. In this case, it is of interest to consider the set of *Pareto-optimal* parameters, that is, the set of parameters that are reasonable in the sense that they are not strictly dominated by any other parameter. We will argue that under correct specification the Pareto-optimal set only contains the true parameters but increases in size in the presence of misspecification. Moreover, we will show how the set of Pareto-optimal parameters can be explicitly calculated on the population level as well as on the sample level in the case of isotonic regression. We also analyze model specification based on the Pareto-optimal set.

The remainder of the chapter is structured as follows. In Section 4.2, we introduce the theoretical concepts we repeatedly use in this work. In Section 4.3, we show that Pareto-optimal parameters characterize correct models. Section 4.4 is devoted to the set of Pareto optimal parameters under misspecification. Moreover, we analyze model specification based on the Pareto-optimal set for some real-world data.

4.2 Loss functions and mixture representations

Statistical models are an important tool that allows us to analyze and predict data from various fields. Often, one has a generic observation (X, Y) consisting of a covariate $X \in \mathbb{R}^p$ and a response variable $Y \in \mathbb{R}$. In this manuscript, we restrict ourselves to the case $p = 1$. In parametric regression, one aims to find a parametric model for the conditional mean

$$g(x) = \mathbb{E}(Y|X = x) = \mathbb{E}(\mathcal{L}(Y|X = x)),$$

slightly abusing notations for the sake of brevity. Let

$$m : \mathbb{R} \times \Theta \to \mathbb{R}, \quad (x, \theta) \to m(x; \theta)$$

be such a parametric model, where Θ is the set of admissible model parameters. The common approach is to determine the optimal parameter θ^0 by

$$\theta^0 = \arg\min_{\theta \in \Theta} \mathbb{E}(m(X; \theta) - Y)^2.$$

Given suitable moment assumptions and that the model m is *correctly specified* for the conditional expectation, that is, $\mathbb{E}(Y|X = x) = m(x; \theta^*)$ almost surely for some $\theta^* \in \mathbb{R}^p$, the above approach yields $\theta^0 = \theta^*$.

In practice however, one merely observes realizations of the tuple (X, Y). We let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). Then, the natural choice to estimate θ^0 is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^{n} (m(X_k; \theta) - Y_k)^2.$$

It is simple to show that under correct specification, estimation of the unknown parameter θ^0 by $\hat{\theta}_n$ yields a consistent estimator for θ^*, subject to moment conditions. However, as mentioned in the introduction, any other Bregman loss function could be chosen in place of the quadratic loss. The class $\mathcal{L}$ of Bregman losses comprises all loss functions of the form

$$L_\phi(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x),$$

where ϕ is a convex function with subgradient ϕ'. Since L_ϕ is the difference between the value of ϕ at point y and the value of its first-order Taylor expansion around point x, L_ϕ takes nonnegative values. Because all Bregman losses are consistent for the mean we have

$$\theta^0 = \arg\min_{\theta\in\Theta} \mathbb{E}L_\phi(m(X;\theta), Y).$$

Then, under correct specification a consistent estimator for θ^* is given by

$$\hat{\theta}_n(\phi) = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} L_\phi(m(X_k;\theta), Y_k).$$

Thus, on the population level and under correct specification minimizing the expected Bregman loss yields the correct parameter θ^* independently of the choice of L_ϕ. When misspecification is present or in the case of finite samples, however, the estimates $\hat{\theta}_n(\phi)$ may vary substantially. This raises questions concerning the choice of loss L_ϕ.

In the introduction, we mentioned how we adapted the concept of *forecast dominance* introduced by Ehm et al. (2016) to obtain a notion of model parameter dominance relative to the class $\mathcal{L}$. When considering finite samples or in the presence of misspecification, however, parameter dominance rarely occurs. This motivates us to consider the set of parameters that are not strictly dominated.

Definition 4.2.1. A parameter θ_1 is said to be *Pareto-optimal* if it is not strictly dominated by any other parameter. In other words, a parameter θ_1 is *Pareto-optimal* if for all $\theta_2 \neq \theta_1$ either there exists $L_\phi \in \mathcal{L}$ such that

$$\mathbb{E}L_\phi(m(X;\theta_1), Y) < \mathbb{E}L_\phi(m(X;\theta_2), Y),$$

or, for all $L_\phi \in \mathcal{L}$, we have

$$\mathbb{E}L_\phi(m(X;\theta_1), Y) = \mathbb{E}L_\phi(m(X;\theta_2), Y).$$

A parameter θ_1 is said to be *weakly Pareto-optimal* if for all $\theta_2 \neq \theta_1$ there exists $L_\phi \in \mathcal{L}$ such that

$$\mathbb{E}L_\phi(m(X;\theta_1), Y) \leq \mathbb{E}L_\phi(m(X;\theta_2), Y).$$

Checking for parameter dominance or Pareto-optimality within the entire class of Bregman losses $\mathcal{L}$ is not directly feasible.

However, Ehm et al. (2016) presented a way out. They showed that any Bregman loss function can be written a mixture of a continuum of *elementary losses*

$$L_\phi(x,y) = \int_{-\infty}^{\infty} S_\eta(x,y)\,\mathrm{d}H_\phi(\eta),$$

where

$$S_\eta(x,y) = (\mathbb{1}\{\eta \le x\} - \mathbb{1}\{\eta \le y\})(\eta - y),$$

and H_ϕ is a nonnegative measure on $\mathbb{R}$ that depends on ϕ. Observe that the parameter η of the elementary losses, S_η, is a scalar parameter η, rather than a convex function ϕ. Thus, analysis of the entire space of η is much more tractable than doing so for ϕ. Ehm et al. (2016) show that forecast dominance with respect to all Bregmann losses is equivalent to forecast dominance with respect to all elementary losses.

Proposition 4.2.2. *A parameter θ_1 is* dominated *by a parameter θ_2 if for all $\eta \in \mathbb{R}$, we have*

$$\mathbb{E}(S_\eta(m(X;\theta_2),Y) \le \mathbb{E}(S_\eta(m(X;\theta_1),Y).$$

A parameter θ_1 is Pareto-optimal *if for all $\theta_2 \neq \theta_1$ either, there exists $\eta \in \mathbb{R}$ such that*

$$\mathbb{E}(S_\eta(m(X;\theta_1),Y) < \mathbb{E}(S_\eta(m(X;\theta_2),Y),$$

or, for all $\eta \in \mathbb{R}$, we have

$$\mathbb{E}(S_\eta(m(X;\theta_1),Y) = \mathbb{E}(S_\eta(m(X;\theta_2),Y).$$

Moreover, parameter θ_1 is weakly Pareto-optimal *if for all $\theta_2 \neq \theta_1$ there exists $\eta \in \mathbb{R}$ such that*

$$\mathbb{E}(S_\eta(m(X;\theta_1),Y) \le \mathbb{E}(S_\eta(m(X;\theta_2),Y),$$

The following properties are immediate consequences of the definitions of parameter dominance and Pareto-optimality.

Proposition 4.2.3. *Assume that θ minimizes all loss functions in $\mathcal{L}$ simultaneously. Then θ is Pareto-optimal and dominates all other parameters.*

Proposition 4.2.4. *Let the parametric forecasting model, $m(x;\theta)$, be point-identified, that is, if $m(x;\theta) = m(x;\theta')$ for (almost) all x then we have $\theta = \theta'$. Then, the unique model parameter is the only Pareto-optimal parameter and dominates all other parameters.*

As we would expect, when the model is correctly specified, then all correct parameters are indeed Pareto-optimal.

Proposition 4.2.5. *Let the parametric forecasting model $m(x;\theta)$ be correctly specified, that is*

$$\mathbb{E}(Y|X) = m(X;\theta^*) \text{ almost surely for some } \theta^* \in \mathbb{R}^p.$$

Then all correct parameters are Pareto-optimal.

Generally, it holds that any correct model parameter dominates any incorrect model parameter. Moreover, any two correct model parameters have equal expected score under any Bregman loss function.

Let us now look at the connection between minimizers of the expected Bregman losses and Pareto-optimal parameters.

Proposition 4.2.6. *Let θ be the unique minimizer of some Bregman loss function L_ϕ. Then θ is Pareto-optimal.*

The following remark elaborates to which extend our results apply to functionals other than the mean functional.

Remark 4.2.7. It is worth mentioning that the previous results do not only apply when the functional T of interest is the mean functional. Pareto-optimal parameters can equivalently be defined for any other class of loss functions. For Propositions 4.2.3 to 4.2.6 to hold, the class of losses $\mathcal{L}$ has to be consistent for functional T. Since, checking for Pareto-optimality is only feasible when the losses L can be written as a mixture of elementary losses, one may consider any elicitable functional T, that is, a functional possessing a strictly consistent loss function, and the class $\mathcal{L}$ of consistent loss functions L for T that can be written as

$$L(x,y) = \int_{-\infty}^{\infty} S_\eta(x,y)\mathrm{d}H(\eta)$$

for some elementary loss functions S_η and a nonnegative measure H on $\mathbb{R}$ associated to L. By Osband's principle, such a mixture representation is always available for sufficiently regular loss functions if the functional T is identifiable with oriented identification function V; see Gneiting (2011); Steinwart et al. (2014); Ziegel (2016b). Then the elementary scores are given by

$$S_\eta(x,y) = (\mathbb{1}\{\eta \le x\} - \mathbb{1}\{\eta \le y\})\, V(\eta, y).$$

The orientation of V ensures that $S_\eta \ge 0$. When T is a τ-expectile or an α-quantile then under mild regularity assumptions the class $\mathcal{L}$ contains all consistent loss functions. For other choices of T class $\mathcal{L}$ may not contain all consistent losses, nevertheless it contains a wide selection of consistent losses. For the mean functional, the identification function is $V(x,y) = (x-y)$, and for the α-quantile, for instance, the identification function is $V(x,y) = \mathbb{1}\{y < x\} - \alpha$.

4.3 Pareto-optimal parameters characterize correct models

Proposition 4.2.3 shows that if a model parameters θ^0 minimizes all losses simultaneously then it is Pareto-optimal. Now, we would like so see whether in this case we have $g(x) = m(x; \theta^0)$. In this section, we show that under some assumptions this is indeed true. If the model is correctly specified and point-identified, then the unique true model parameter is the only Pareto-optimal parameter.

Let (X, Y) be a pair of real valued random variables. Let $\mathcal{P}$ denote the class of all conditional distributions $\mathcal{L}(Y|X = x)$. Let $\mathbb{E} : \mathcal{P} \to \mathbb{R}$ be the mean functional on $\mathcal{P}$ with oriented identification function $V(x, y) = x - y$, where $P \subset \mathbb{R}$ denotes the set of admissible predictions. Let $\mathcal{F} = \{m(\cdot\,; \theta) \colon \mathbb{R} \to \mathbb{R} \mid \theta \in \Theta\}$ be a family of parametric models for $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto \mathbb{E}(Y|X = x)$.

We would like to show that if $\theta(\phi) = \theta^0$ for all ϕ, where $\theta(\phi) = \arg\min_\theta L_\phi(m(X; \theta), Y)$, then we have $g(\cdot) = m(\cdot\,; \theta^0)$. Note that this conjecture is generally false, for example, when the model is not rich enough. Indeed, suppose that Θ only has one element $\Theta = \{\theta^*\}$ then we always trivially have $\theta(\phi) = \theta^*$. However, there is no good reason why $g = m(\cdot\,; \theta^*)$ should hold.

Therefore, some assumptions are needed to show a suitable result.

Definition 4.3.1. Let $U \subseteq \mathbb{R}$ be an open set and let I be an open interval containing 0. A function $h : I \times U \to U, (a, u) \mapsto h_a(u)$ is called *admissible dominator function* if the following conditions hold.

1. h_a is increasing for each $a \in I$.
2. $h_0 \colon U \to U$ is the identity.
3. For each $a \in I$, there exists an $a' \in I$ such that $h_a^{-1} = h_{a'}$.
4. $h_a(u)$ is continuous in a at $a = 0$ for any $u \in U$.

Assumption 4.3.2. Let $U \subseteq \mathbb{R}$ be an open set that contains the range of values of the models $m(\cdot\,; \theta) \in \mathcal{F}$, that is $\bigcup_\theta \operatorname{range}(m(\cdot\,; \theta)) \subseteq U \subseteq \mathbb{R}$. Let I be an open interval containing 0 and let h be an admissible dominator function on $I \times U$. Suppose that for any model $m(\cdot\,; \theta) \in \mathcal{F}$ and $a \in I$, we also have $h_a(m(\cdot\,; \theta)) \in \mathcal{F}$.

Suppose that the range of the values of the models $m(\cdot\,; \theta) \in \mathcal{F}$ is $\mathbb{R}$, which is the natural range for linear models for the mean. Then an admissible dominator function h would be

$$h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (a, u) \mapsto u + a.$$

Assumption 4.3.2 is then the requirement that to any linear model in $\mathcal{F}$ we can add an arbitrary intercept value. If we want to consider a class of linear models $\mathcal{F}$ with intercept

forced to zero, but we allow for all possible slopes, we can consider the admissible dominator function

$$h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, (a, u) \mapsto e^a u.$$

This latter admissible dominator function is also interesting for models whose range is $(0, \infty)$ such as volatility models.

Theorem 4.3.3. *Suppose that $\theta^* \in \Theta$ dominates all other parameters i.e. $\theta(\phi) = \theta^*$ for all ϕ and let assumption 4.3.2 hold. Then $m(X; \theta^*) = \mathbb{E}(Y|m(X; \theta^*))$ almost surely.*

The proof can be found in the appendix. Intuitively one might rather expect a conclusion like $\mathbb{E}(Y|X) = m(X; \theta^*)$ almost surely. However, if $\sigma(m(X; \theta^*)) \supseteq \sigma(g(X))$, then Theorem 4.3.3 implies that $\mathbb{E}(Y|X) = m(X; \theta^*)$ almost surely. The following example shows this for the case where we consider linear models $\theta_0 + \theta_1 x$, $x \in \mathbb{R}$, and $\theta_1^* \neq 0$.

Example 4.3.4. Suppose that $Y = g(X) + \epsilon$ with X and ϵ independent and $\mathbb{E}(\epsilon) = 0$. Then $\mathbb{E}(Y|X = x) = g(x)$. Let $m(x; \theta) = \theta_0 + \theta_1 x$ and suppose there is a $\theta^* \in \mathbb{R}^2$ that dominates all other parameters. Then if $\theta_1^* \neq 0$, we have $\sigma(X) = \sigma(m(X; \theta^*))$ and

$$\mathbb{E}(Y|X = x) = g(x) = m(x; \theta^*).$$

On the other hand, if $\theta_1^* = 0$, then $\sigma(X) \neq \sigma(m(X; \theta^*)) = \{\Omega, \emptyset\}$ and

$$\mathbb{E}(Y|X = x) = g(x) \text{ and } \mathbb{E}(Y|m(X; \theta^*)) = \mathbb{E}(Y).$$

An open question that remains is whether it can happen that $\theta(\phi) = \theta^*$, for all ϕ, but $m(X; \theta^*) \neq g(X)$ almost surely. So far, we have not been able to construct a non-trivial example. By trivial, we mean that the models in $\mathcal{F}$ are simply missing a covariate, that is, for $X = (X_1, \ldots, X_k)$ we have $g(x) = \mathbb{E}(Y|X = x)$ but $m(x; \theta) = \mathbb{E}(Y|(X_1, \ldots, X_{k-1}) = (x_1, \ldots, x_{k-1}))$.

Remark 4.3.5. Similar to Section 4.2 the results in this section can easily be generalized to other elicitable functionals T possessing an oriented identification function V. A remark on the additional assumptions needed for the proof of Theorem 4.3.3 can be found in the appendix.

4.4 Pareto-optimal parameters inform about misspecification

In the last section, we have seen that a unique Pareto-optimal parameter implies correct model specification. And from Section 4.2 we know that correctly a correctly specified model has a unique Pareto-optimal parameter. In this section, we turn out attention to the set of Pareto-optimal parameters under model misspecification. To this end, we consider linear regression models.

As mentioned in the introduction, Buja et al. (2019) argue convincingly that the presence of nonlinearity can conspire with the randomness of the regressors to create sampling variability in the slope and intercept estimates. They consider $m(X, \theta)$ to be a linear model and introduce the following decomposition

$$Y = m(X; \theta^0) + (g(X) - m(X; \theta^0)) + (Y - g(X)),$$

where $\epsilon := (Y - g(X))$ is the noise, and $n(X) := (g(X) - m(X; \theta^0))$ is the nonlinearity term. In linear regression, one assumes that $n(X) = 0$ almost surely and ϵ is independent of X and has expectation zero. When applying this decomposition, we obtain

$$\begin{aligned}
\arg\min_\theta \mathbb{E} S_\eta(m(X;\theta), Y) &= \arg\min_\theta \mathbb{E}\mathbb{1}\{\eta \leq m(X;\theta)\}(\eta - Y) \\
&= \arg\min_\theta \Big(\mathbb{E}\mathbb{1}\{\eta \leq m(X;\theta)\}\big(\eta - m(X;\theta^0)\big) \\
&\qquad - \mathbb{E}\mathbb{1}\{\eta \leq m(X;\theta)\} n(X) \\
&\qquad - \mathbb{E}\mathbb{1}\{\eta \leq m(X;\theta)\}\epsilon \Big)
\end{aligned}$$

If it holds that $n(X) = 0$ almost surely and ϵ is independent of X and has expectation zero, then the second and the third term cancel and the arg min contains θ^0 for all $\eta \in \mathbb{R}$. However, one can see directly in the above decomposition that when nonlinearity, and therefore misspecification, is present, i.e. $n(X) \neq 0$ almost surely, then different values of $\eta \in \mathbb{R}$ might lead to different parameters with larger expected elementary loss than θ^0 even in the absence of error. In a mildly misspecified case, the second term might not yet have too much weight in the minimization, so that for most η the minimization still yields a parameter with equal expected elementary loss as θ^0. However, the stronger the misspecification becomes the more weight is put on the second term so that the size of the set of Pareto-optimal parameters increases with the degree of misspecification. One of the issues formulated by Buja et al. (2019) is that over a narrow covariate range a model has a better change of appearing well-specified. This can also be seen in the above decomposition, for a narrow range the second term may not have a huge weight since linear approximations generally work better locally.

These rough considerations arouse interest in understanding what the Pareto-optimal set actually looks like. Under the assumption that g is isotonic, we show how the set of Pareto-optimal parameters for a linear model can be calculated explicitly.

4.4.1 Pareto-optimal parameters in isotonic regression problems

The following auxiliary result shows that in the case of an isotonic regression function g and a simple linear model $m(\theta; x) = \theta_0 + \theta_1 x$ for the latter only minimizers of some elementary loss can be Pareto-optimal.

Proposition 4.4.1. *Let $Y, X \in \mathbb{R}$ and suppose that the support $\operatorname{supp}(X)$ is a (possibly unbounded) interval. Assume that the regression function $g(x) = \mathbb{E}(Y \mid X = x)$ is strictly increasing and differentiable. Consider a simple linear model $m(\theta; x) = \theta_0 + \theta_1 x$ for $g(x)$, where $\theta = (\theta_0, \theta_1) \in \Theta = \mathbb{R} \times (0, \infty)$.*

(a) Assume that

$$\liminf_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) \neq 0, \qquad \liminf_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) \neq 0,$$
$$\limsup_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) \neq 0, \qquad \limsup_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) \neq 0,$$

for all $\theta \in \Theta$. Then all Pareto-optimal parameters are minimizers of some elementary loss.

(b) If there exists a parameter $\theta \in \Theta$ such that

$$\liminf_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) = 0,$$

then θ is uniquely determined. Similarly, if $\theta \in \Theta$ is such that

$$\liminf_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) = 0, \quad \limsup_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) = 0,$$
$$\text{or } \limsup_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0) - g^{-1}(\eta)) = 0,$$

respectively, then θ is uniquely determined.

The argument behind this proposition is that the model $m(x;\theta)$ minimizes $\mathbb{E}S_\eta$ if, and only if, the η-superlevel set of $g(x)$ and $m(x;\theta)$ coincide. If the model $m(x;\theta')$ is not a minimizer, then this means, that the graphs of $m(x;\theta')$ and $g(x)$ do not touch. By adjusting the intercept, we can obtain a model $m(x;\theta)$ so that its graph touches the graph of $g(x)$. This new model is not worse than the original model, but outperforms the original model at the point of contact. Therefore, the original model parameter cannot be Pareto-optimal. A rigorous proof is given in the appendix.

The following theorem provides a complete characterization of the set of Pareto-optimal parameters in the case of an isotonic regression function g.

Theorem 4.4.2. *Let $Y, X \in \mathbb{R}$ and suppose that the support $\operatorname{supp}(X)$ is a (possibly unbounded) interval. Assume that $g(x) = \mathbb{E}(Y \mid X = x)$, where g is strictly increasing and differentiable. Moreover, consider the parametric model*

$$m : \mathbb{R} \times \Theta \to \mathbb{R}, \quad (x, \theta) \mapsto m(x;\theta) = \theta_0 + \theta_1 x,$$

where $\Theta = \mathbb{R} \times (0, \infty)$.

(a) Assume that

$$\liminf_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))\neq 0, \qquad \liminf_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))\neq 0,$$
$$\limsup_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))\neq 0, \qquad \limsup_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))\neq 0,$$

for all $\theta=(\theta_0,\theta_1)$. Then, the set of Pareto-optimal parameters consists of all parameters θ of the form

$$\theta_0 = g(x^0)-g'(x^0)x^0, \quad \theta_1 = g'(x^0), \quad x^0\in\operatorname{supp}(X), \tag{4.1}$$

and all parameters $\theta\in\Theta$ that solve the following system of equations

$$\begin{aligned} g(x^1) &= \theta_0+\theta_1 x^1 \\ g(x^2) &= \theta_0+\theta_1 x^2 \end{aligned} \tag{4.2}$$

for $x^1, x^2\in\operatorname{supp}(X)$, $x^1\neq x^2$.

(b) Assume that there exist parameters $\theta\in\Theta$ such that $1\leq r\leq 4$ of the following equalities are fulfilled

$$\liminf_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))=0, \qquad \liminf_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))=0,$$
$$\limsup_{\eta\to\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))=0, \qquad \limsup_{\eta\to-\infty}(\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta))=0.$$

Then, there are possibly r additional Pareto-optimal parameters.

The proof of Theorem 4.4.2 (a) relies on the following argument. Whenever the line given by $m(x;\theta)$ is not a tangent or chord of g, that is a line that intersects g twice, then slightly rotating the line until it touches g a second time yields a model with dominating model parameter. When $m(x;\theta)$ corresponds to a tangent or a chord of g the rotation of the line will not improve or but rather worsen the expected score at the point of contact. A rigorous proof can be found in the appendix. Part (b) of Theorem 4.4.2 corresponds to the case, where g has a linear asymptote.

Recall that Buja et al. (2019) mentioned that the presence of nonlinearity in g affects the parameter estimates even in the complete absence of error. In this case, the optimal parameters depend on the sample. It turns out that the set of Pareto-optimal parameters captures exactly this variability as the lines generated from the Pareto-optimal parameters correspond to all tangents and chords of g.

Setting $x_0=g^{-1}(\eta)$ we can alternatively write the Pareto-optimal set as all parameters given by

$$\theta_0 = \eta-g^{-1}(\eta)g'(g^{-1}(\eta)), \quad \theta_1 = g'(g^{-1}(\eta)), \quad \eta\in\operatorname{image}(g) \tag{4.3}$$

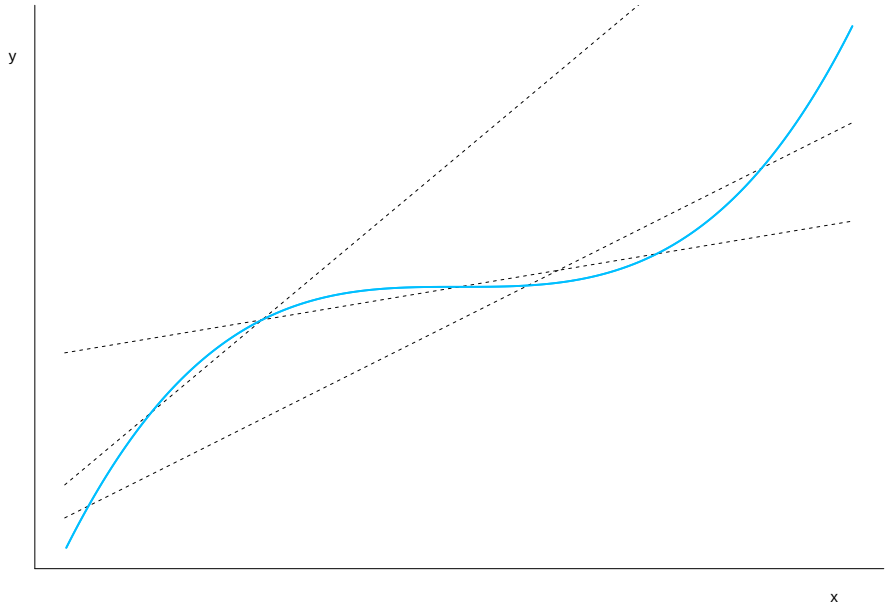

Figure 4.1: For $g(x) = x^3$ three linear models with Pareto-optimal parameters are displayed.

and all parameters θ such that the function

$$\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$$

has at least two zeros. Figure 4.1 contains a sample of Pareto-optimal fits to an increasing function g.

The following example shows that under correct specification the result in Theorem 4.4.2 does indeed yield the true parameters.

Example 4.4.3. If our model is correctly specified i.e. $g(x) = a + bx$ with $b > 0$ then the Pareto-optimal parameters are the true parameters. Indeed,

$$\theta_1 = g'\left(g^{-1}(\eta)\right) = b \quad \text{and} \quad \theta_0 = \eta - \frac{\eta - a}{b} \cdot b = a.$$

The next example shows how the Pareto-optimal set may look like under misspecification.

Example 4.4.4. (a) Let $g(x) = e^x/(1+e^x)$. Then, $g'(\eta) = e^\eta/(1+e^\eta)^2$ and $g^{-1}(\eta) = \log(\eta) - \log(1-\eta)$. Hence, the set of Pareto-optimal parameters is given by

$$\theta_0 = \eta\left(1 - (1-\eta)\log\left(\frac{\eta}{1-\eta}\right)\right) \text{ and } \theta_1 = \eta(1-\eta) \text{ for } \eta \in [0,1],$$

and all parameters $\theta \in \Theta$ such that

$$\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$$

has at least two zeros. Figure 4.2 shows the Pareto-optimal set for this function g.

(b) Let $g_2(x) = x^3$ then $g_1'(\eta) = 3\eta^2$ and $g_1^{-1}(\eta) = \eta^{1/3}$. Hence, the set of Pareto-optimal parameters is given by

$$\theta_0 = -2\eta \text{ and } \theta_1 = 3\eta^{2/3} \text{ for } \eta \in \mathbb{R},$$

and all parameters $\theta \in \Theta$ such that

$$\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$$

has at least two zeros, i.e. all parameters $\theta \in \Theta$ with $\theta_1^3/\theta_0^2 > 27/4$. Figure 4.3 contains the Pareto-optimal set for this case.

(c) Consider $g(x) = 1 - e^{-x} + x$. This function is strictly increasing, differentiable and concave. The Pareto-optimal parameters according to (4.3) are given by

$$\theta_0 = \eta - \left(1 + W(e^{1-\eta})\right)\left(\eta + W(e^{1-\eta}) - 1\right)$$

and

$$\theta_1 = 1 + W(e^{1-\eta})$$

for $\eta \in \mathbb{R}$, where W is the Lambert-W function (Corless et al., 1996). Moreover, the function $\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$ has two zeros for $\theta \in \Theta$ with $\theta_1 > 1$ and

$$\theta_0 < 2 + \theta_1 - \log(\theta_1 - 1)(1 - \theta_1).$$

Figure 4.4 shows the Pareto-optimal set.

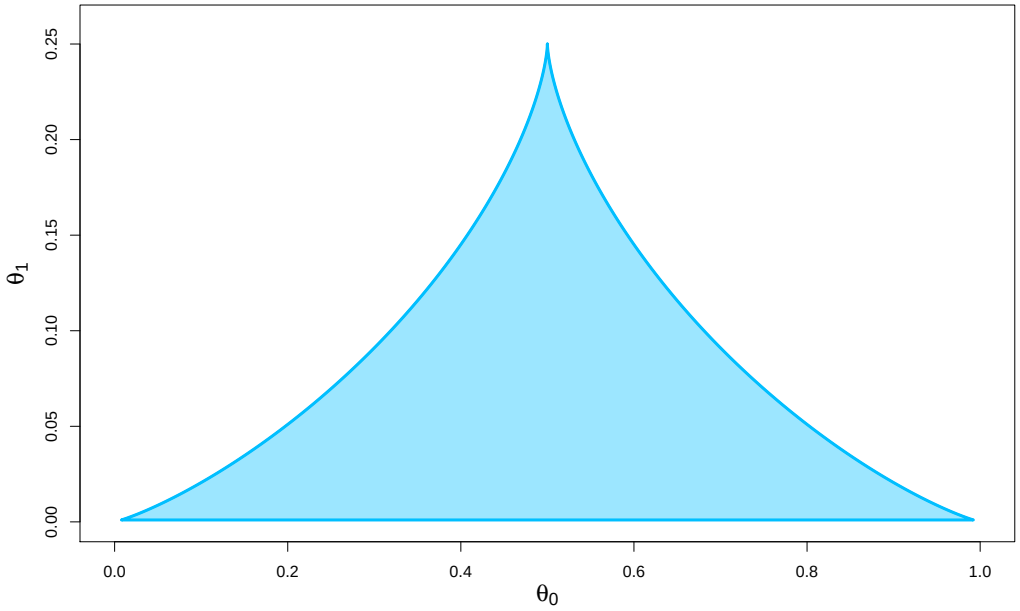

Figure 4.2: The set of Pareto-optimal parameters for $g_1(x) = e^x/(1+e^x)$ is drawn in blue.

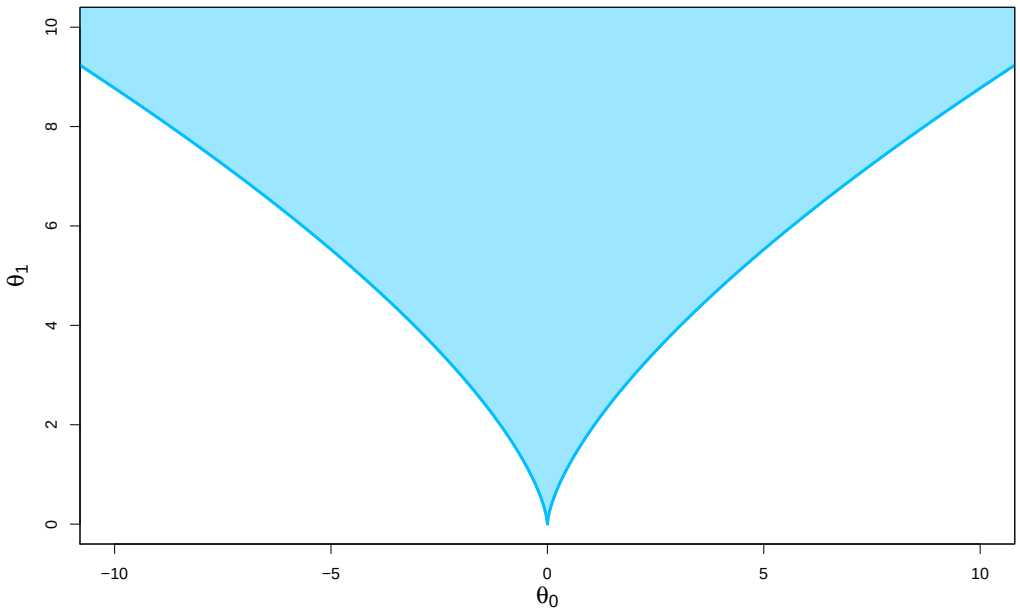

Figure 4.3: In blue the set of Pareto-optimal parameters for $g_2(x) = x^3$ is displayed.

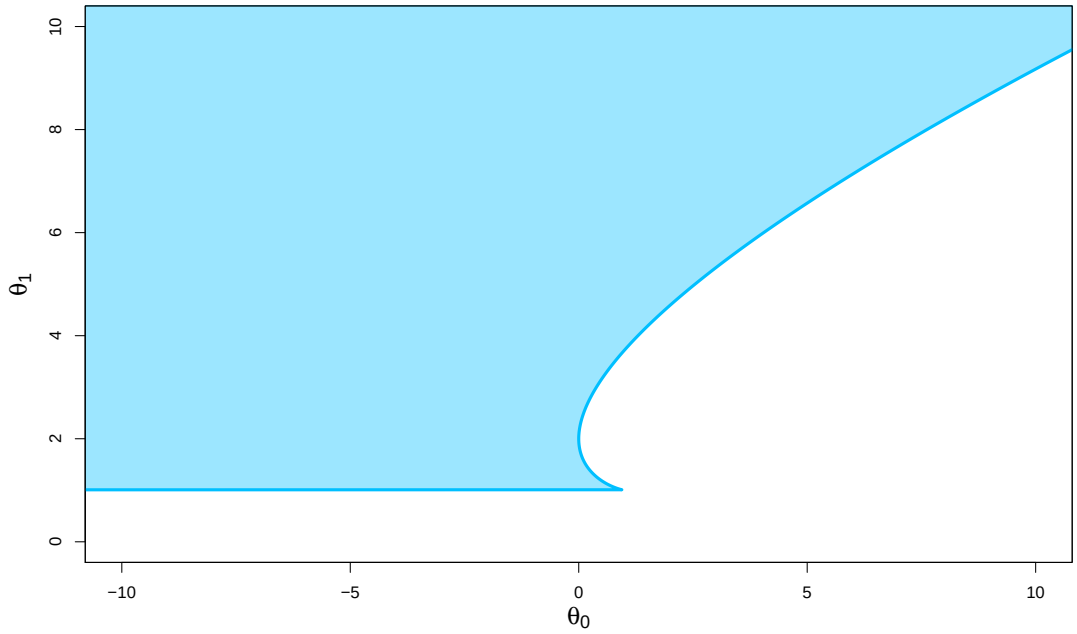

Figure 4.4: In blue the set of Pareto-optimal parameters for $g(x) = 1 - e^{-x} + x$ is drawn.

4.4.2 Calculation on the sample level

In practice, the true function g is unknown, otherwise we would not be forced to estimate g from the data. Thus, it is necessary to be able to determine the Pareto-optimal parameters on the sample level. For the setting of Theorem 4.4.2 the following result shows how the Pareto optimal set can be calculated on the sample level.

Proposition 4.4.5. *Let random variables $X, Y \in \mathbb{R}$ be such that $g(x) = \mathbb{E}(Y|X = x)$ is increasing. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent observations. Consider the linear model*

$$m : \mathbb{R} \times \Theta \to \mathbb{R}, \quad (x, \theta) \mapsto m(x; \theta) = \theta_0 + \theta_1 x,$$

where $\Theta = \mathbb{R} \times (0, \infty)$. Then set of Pareto-optimal parameters consists of all parameters $\theta = (\theta_0, \theta_1)$ such that

$$\{x : m(X; \theta) \geq \eta\} = \{x : \hat{g}_{PAV}(x) \geq \eta\}$$

for at least two different superlevel sets of $\hat{g}_{PAV}$ that are not equal to the empty set or $\{X_1, \ldots, X_n\}$, where $\hat{g}_{PAV}$ denotes the optimal isotonic fit as defined in Barlow et al. (1972).

For a specific sample Figure 4.5 shows the optimal isotonic fit to the data as well as three linear fits with Pareto-optimal parameters.

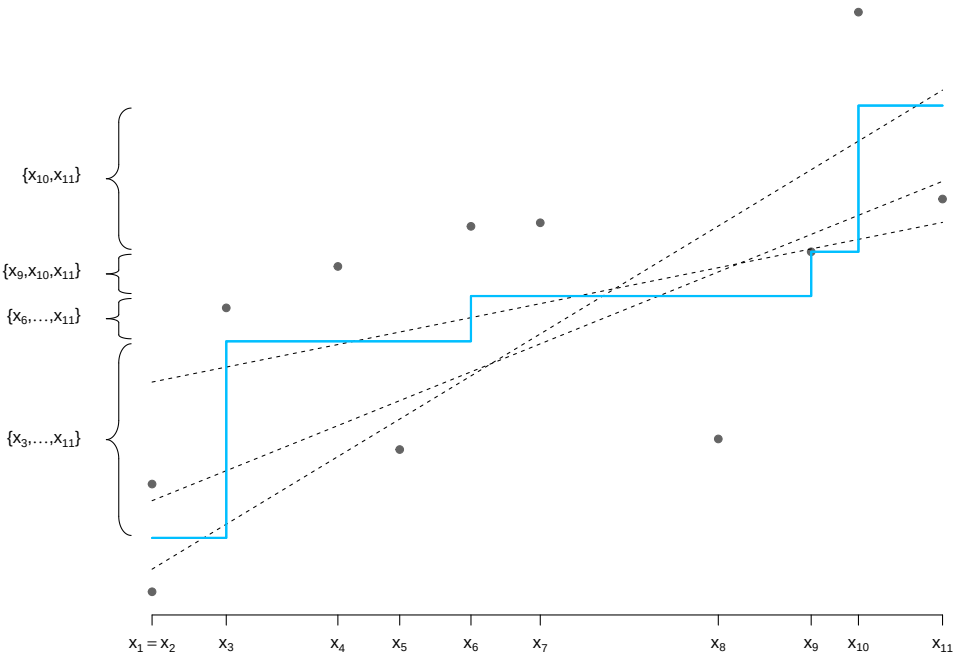

Figure 4.5: For a specific sample of 11 data points the PAVA-fit $\hat{g}_{PAV}$ is drawn in blue. Moreover, the optimal superlevel sets are displayed. Three linear models with Pareto-optimal parameters are drawn with dashed lines.

Proposition 4.4.5 provides us with the possibility to explicitly calculate the set of Pareto-optimal parameters for a sample. In order to do so, we discretize the space of η's. Because the optimal isotonic fit $\hat{g}_{PAV}$ is rather rough the resulting Pareto-optimal sets can comprise some outliers. This can occur when for close values η_1, η_2 the superlevel sets of $\hat{g}_{PAV}$ are different but rather similar, in that the set difference contains few x. In this case, it may well be that the superlevel sets of the true regression function g would actually be equal. Therefore, we removed the parameters that correspond to lines going through two superlevel sets $\{x : \hat{g}_{PAV}(x) \geq \eta_1\}$ and $\{x : \hat{g}_{PAV}(x) \geq \eta_2\}$ with $|\eta_2 - \eta_1| < \delta$. The reason for this is that these lines tend to be very steep due to the local roughness of the optimal isotonic fit even if entire fit is rather flat. What a reasonable choice of smoothing parameter δ is depends on the scale of g. In the figures below, we chose, $\delta = 10^{-1} \max_{\eta_1,\eta_2} |\eta_2 - \eta_1|$ and $\delta = 5^{-1} \max_{\eta_1,\eta_2} |\eta_2 - \eta_1|$. Figures 4.6 and 4.7 illustrate the effect of the smoothing on the Pareto-optimal set. Figure 4.6 contains the original as well as some smoothed versions of the Pareto-optimal set under misspecification. Figure 4.7 contains the original as well as some smoothed versions of the Pareto-optimal set under correct specification.

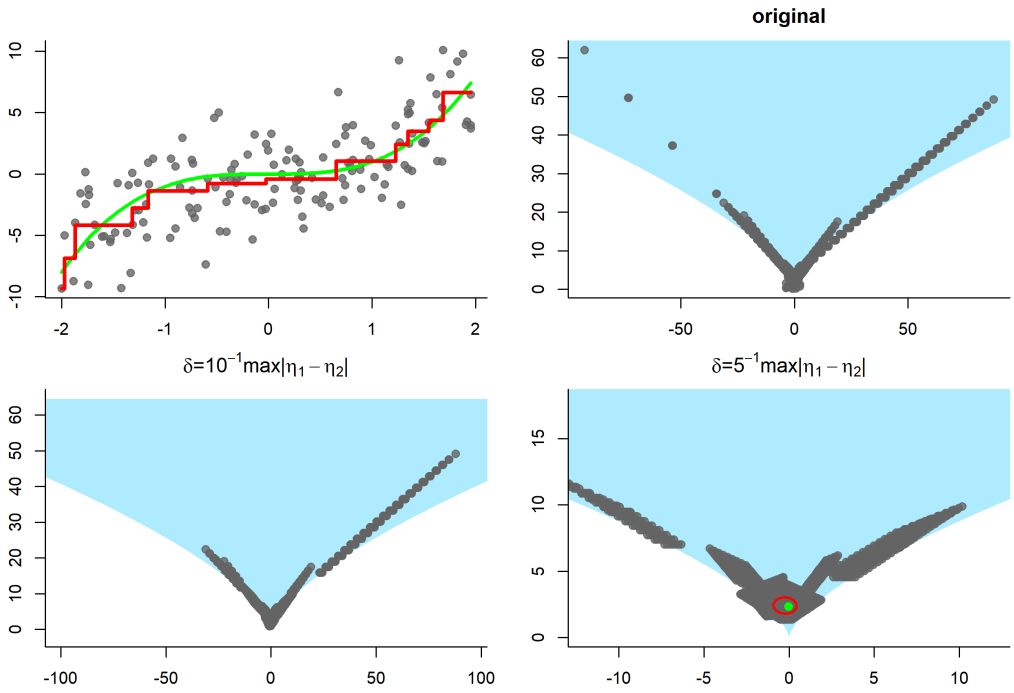

Figure 4.6: On the top left, for a sample of size 150, the true function $g(x) = x^3$ is drawn in green along with the PAVA-fit $\hat{g}_{PAV}$ in red. On the top right, the empirical Pareto-optimal set for this sample is displayed (gray). The blue set is the Pareto-optimal set with respect to the true function g. On the bottom smoothed versions of the Pareto-optimal set are drawn for two δ whereof the picture on the bottom right has smaller scale. The green point corresponds to the OLS estimate and the red line is the corresponding 95%-confidence ellipsoid.

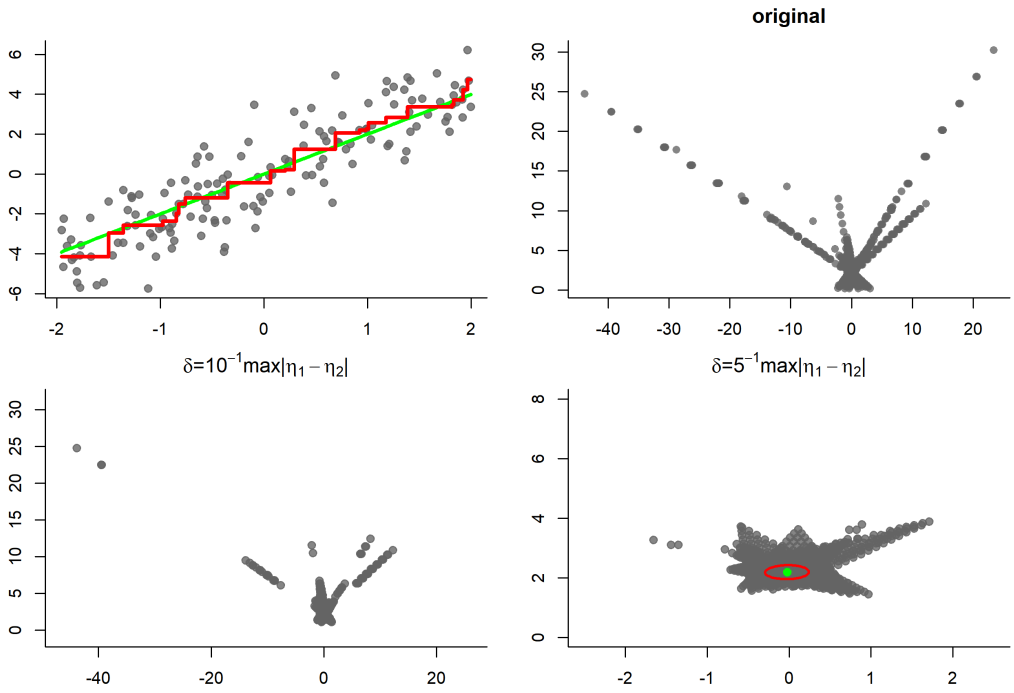

Figure 4.7: On the top left, for a sample of size 150, the true function $g(x) = 2x$ is drawn in green along with the PAVA-fit $\hat{g}_{PAV}$ in red. On the top right, the empirical Pareto-optimal set for this sample is displayed (gray). The green point corresponds to the OLS estimate while the red line is the 95%-confidence interval for the OLS-estimate. On the bottom smoothed versions of the empirical Pareto-optimal set are drawn for two δ whereof the picture on the bottom right has a smaller scale.

4.4.3 Evaluation of two data examples

Finally, we use our methodology to assess the model misspecification in two data examples. We consider two data sets published in Franses (1998). The first data set of size 80 contains the quarterly Consumer Price Index of Argentina from 1970 to 1989. In a first step, the log price difference $\log(y_t) - \log(y_{t-1})$ is calculated. Then one aims to predict $\log(y_t) - \log(y_{t-1})$ from $\log(y_{t-1}) - \log(y_{t-2})$. This approach is called an auto-regressive model of order one and is denoted by $AR(1)$. The model equation of an $AR(1)$ model is $x_t = \theta_0 + \theta_1 x_{t-1} + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, 1)$. In the aforementioned context we have $x_t = \log(y_t) - \log(y_{t-1})$. On page 52, Franses (1998) concludes by analyzing estimated auto-correlation function patterns and estimated partial auto-correlation function patterns that an $AR(1)$ model is indeed adequate. We analyze the set of Pareto-optimal parameters for this data example. The first column in Figure 4.8 contains the results for this data set. The corresponding OLS estimate is given by $\hat{\theta}_{OLS} = (0.13, 0.56)$. When looking at the smoothed set of Pareto-optimal parameters on the bottom left of Figure 4.8, we arrive at the same conclusion as Franses (1998). However, it should be mentioned that the last three observations are outliers. The rapid increase in inflation is due to a new economic plan approved in 1988 during the presidency of Raúl Alfonsín. This can be seen in the time series on the top left of Figure 4.8. If we would cut the time series after 1988, we obtain the results contained in the second column of Figure 4.8. These new results, however, hint that a linear fit might actually not be as suitable.

As a second example, we consider a data set of size 120 containing the seasonally adjusted quarterly unemployment rate y_t^{adj} in Germany from 1962 to 1991 (Franses, 1998). Again, we assume an $AR(1)$-model for y_t^{adj}. The OLS estimate is given by $\hat{\theta}_{OLS} = (0.079, 0.993)$. The corresponding set of Pareto-optimal parameters in Figure 4.9 is rather large in size compared to the scale we have seen in the previous data example or in the simulation example in Figure 4.7. This observation hints that an $AR(1)$-model is probably misspecified for this data set which concurs with the findings in Franses (1998).

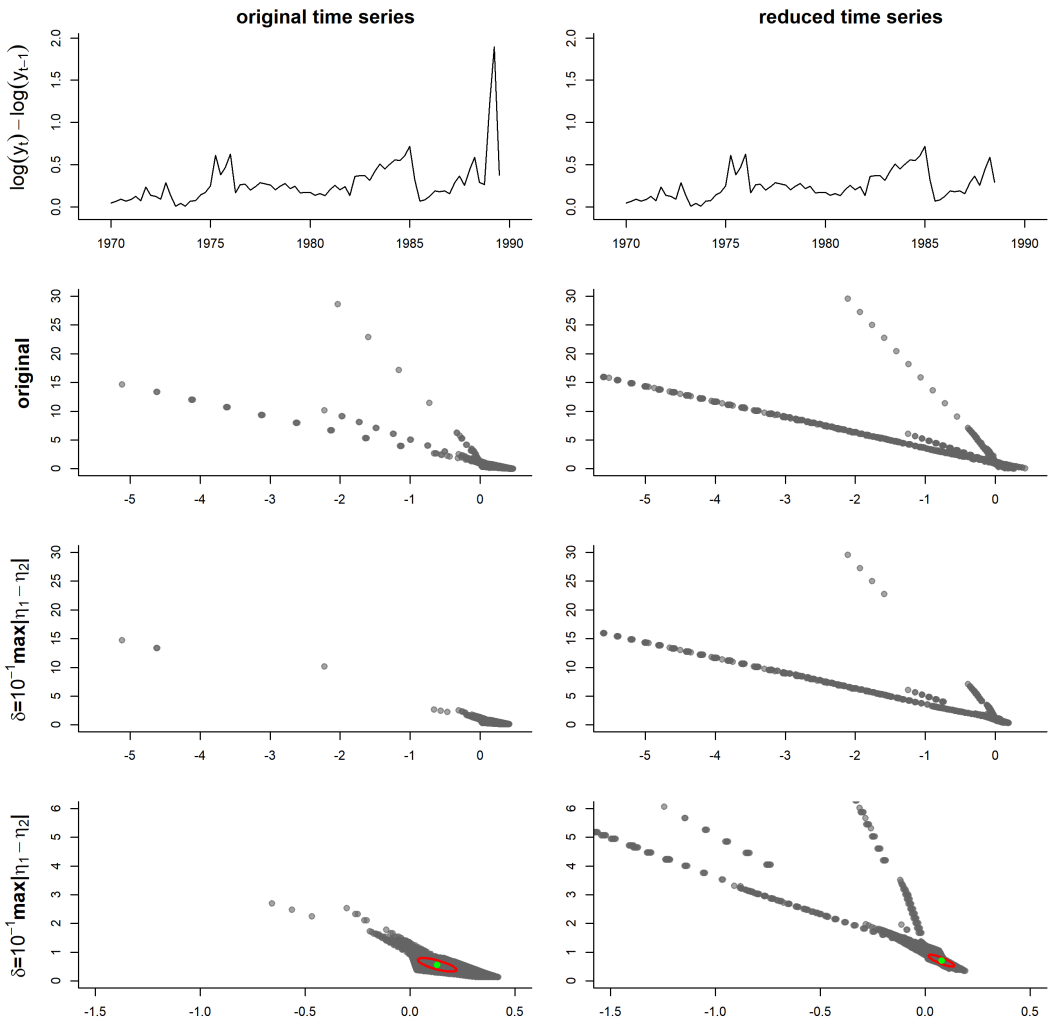

Figure 4.8: The left column contains the results for the original time series of the log price differences $\log(y_t) - \log(y_{t-1})$, the right column contains the results for the same time series but cut off after 1988 and thereby removing the outliers. In both columns, the top picture displays the respective time series. The second row contains the respective sets of Pareto-optimal parameters. The third row displays a smoothed version of the Pareto-optimal set in the second row. Finally, the bottom row contains a close up view of the smoothed Pareto-optimal set. The OLS-estimates are drawn in green while the 95%-confidence ellipsoids for the respective OLS-estimates are drawn in red.

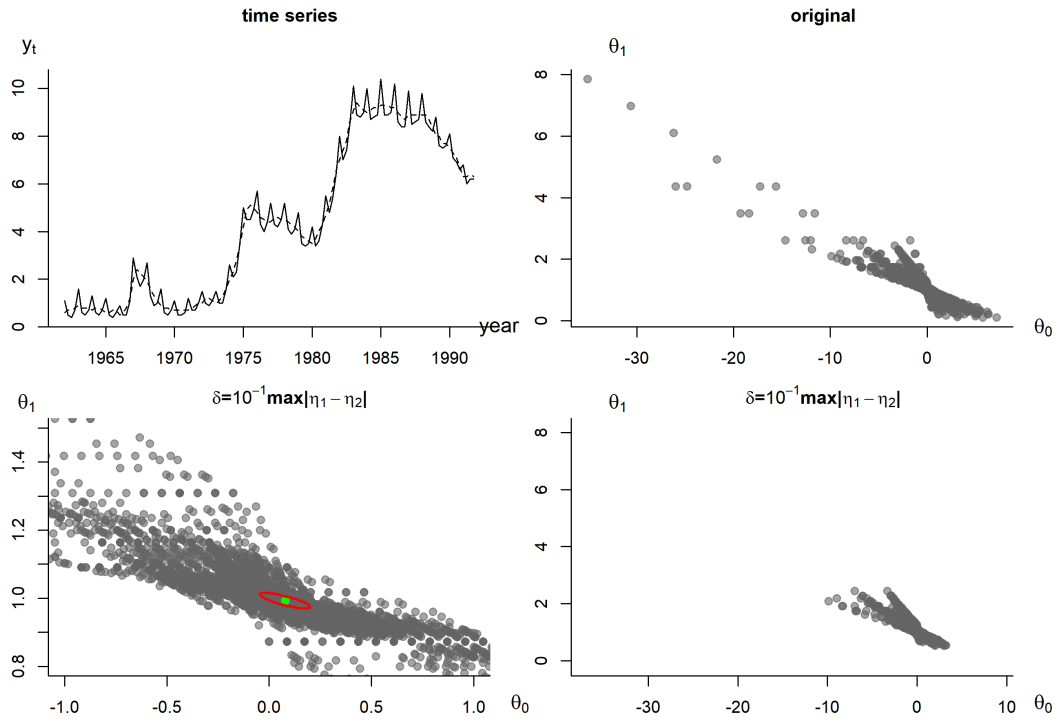

Figure 4.9: On the top left the time series of the quarterly unemployment rate y_t is displayed in black. The dotted line corresponds to the seasonally adjusted unemployment rate y_t^{adj}. On the top right the set of Pareto-optimal parameters. On the bottom left a smoothed version (right) as well as its close-up (left) are drawn. The OLS-estimate is drawn in green whereas the 95%-confidence ellipsoid for the OLS-estimate is drawn in red.

4.A Proofs

Lemma 4.A.1. *Suppose that Assumption 4.3.2 holds. If $\theta(\phi) = \theta^*$ for all ϕ, then for all $a \in I$, $\theta \in U$, we have*

$$\begin{aligned}&\mathbb{E}\mathbb{1}\{m(X;\theta^*) \in [\theta, h_a(\theta))\}(h_a(\theta) - Y) \geq 0,\\ &\mathbb{E}\mathbb{1}\{m(X;\theta^*) \in [\theta, h_a(\theta))\}(\theta - Y) \leq 0.\end{aligned}$$

Proof of Lemma 4.A.1. As $\theta(\phi) = \theta^*$ for all ϕ, we have for any $a \in I$ and $\theta \in U$

$$\mathbb{E}(\mathbb{1}\{\theta \leq m(X;\theta^*)\} - \mathbb{1}\{\theta \leq h_a(m(X;\theta^*))\})V(\theta, Y) \leq 0.$$

This is equivalent to

$$\mathbb{E}\mathbb{1}\{m(X;\theta^*) \in [\theta, h_a^{-1}(\theta))\}V(\theta, Y) \leq 0. \tag{4.4}$$

By property 3 of Definition 4.3.1, equation (4.4) also holds for h_a^{-1} replaced with h_a. On the other hand, we also obtain

$$\mathbb{E}(\mathbb{1}\{h_a(\theta) \leq m(X;\theta^*)\} - \mathbb{1}\{h_a(\theta) \leq h_a(m(X;\theta^*))\})V(h_a(\theta), Y) \leq 0,$$

which is equivalent to

$$\mathbb{E}\mathbb{1}\{m(X;\theta^*) \in [\theta, h_a(\theta))\}V(h_a(\theta), Y) \geq 0.$$

□

Proof of Theorem 4.3.3. Recall that we consider $V(x, y) = (x - y)$. Let $\lambda_{n,k} = k2^{-n}$ for $k \in \mathbb{Z}$, $n \in \mathbb{N} \cup \{0\}$. Define

$$\begin{aligned}Z_n &:= \sum_{k\in\mathbb{Z}} \lambda_{n,k}\mathbb{1}\{m(X;\theta^*) \in [\lambda_{n,k}, \lambda_{n,k+1})\},\\ W_n &:= \sum_{k\in\mathbb{Z}} \lambda_{n,k+1}\mathbb{1}\{m(X;\theta^*) \in [\lambda_{n,k}, \lambda_{n,k+1})\},\end{aligned}$$

and

$$\mathcal{A}_n := \sigma(\{m(X;\theta^*) \in [\lambda_{n,k}, \lambda_{n,k+1})\}, k \in \mathbb{Z}).$$

We have $Z_n \leq Z_{n+1}$, $W_n \geq W_{n+1}$ and thus by the monotonicity of $V(\cdot, y)$, we obtain $V(Z_n, Y) \leq V(Z_{n+1}, Y)$ and $V(W_n, Y) \geq V(W_{n+1}, Y)$. Furthermore, $Z_n \uparrow m(X;\theta^*)$ and $W_n \downarrow m(X;\theta^*)$ as $n \to \infty$ and

$$\begin{aligned}&\mathbb{E}V(Z_0, Y) = \mathbb{E}V(\lfloor m(X;\theta^*)\rfloor, Y) \leq \mathbb{E}V(m(X;\theta^*), Y) < \infty\\ &\mathbb{E}V(W_0, Y) = \mathbb{E}V(\lfloor m(X;\theta^*) + 1\rfloor, Y) \leq \mathbb{E}V(m(X;\theta^*), Y) + 1 < \infty.\end{aligned} \tag{4.5}$$

Note that $\mathcal{A}_n \subset \mathcal{A}_{n+1}$ because

$$\begin{aligned}\{m(X;\theta^*) \in [\lambda_{n,k}, \lambda_{n,k+1})\} = &\{m(X;\theta^*) \in [\lambda_{n+1,2k}, \lambda_{n+1,2k+1})\} \\ &\cup \{m(X;\theta^*) \in [\lambda_{n+1,2k+1}, \lambda_{n+1,2k+2})\}.\end{aligned}$$

We define $\bar{Z}_n := \mathbb{E}(V(Z_n, Y)|\mathcal{A}_n)$ and $\bar{W}_n := \mathbb{E}(V(W_n, Y)|\mathcal{A}_n)$. For n large enough, Lemma 4.A.1 implies that $\bar{Z}_n \leq 0$ almost surely and that $\bar{W}_n \geq 0$ almost surely as the generator of $\mathcal{A}_n$ consists of disjoint sets. Furthermore,

$$\begin{aligned}\mathbb{E}(\bar{Z}_{n+1}|\mathcal{A}_n) &= \mathbb{E}(\mathbb{E}(V(Z_{n+1}, Y)|\mathcal{A}_{n+1})|\mathcal{A}_n) \\ &= \mathbb{E}(V(Z_{n+1}, Y)|\mathcal{A}_n) \\ &\geq \mathbb{E}(V(Z_n, Y)|\mathcal{A}_n) = \bar{Z}_n \text{ almost surely,}\end{aligned}$$

and, analogously, $\mathbb{E}(\bar{W}_{n+1}|\mathcal{A}_n) \leq \bar{W}_n$ almost surely. Therefore $(\bar{Z}_n)_n$ is a non-positive sub-martingale and $(\bar{W}_n)_n$ is a non-negative super-martingale with respect to $(\mathcal{A}_n)_n$. Corollary 1 in (Bauer, 1974, 60.2) implies that there exists $\bar{Z}_\infty \leq 0$ and $\bar{W}_\infty \geq 0$ integrable such that $\bar{Z}_n \to \bar{Z}_\infty$, $\bar{W}_n \to \bar{W}_\infty$ almost surely as $n \to \infty$ and $\mathbb{E}(\bar{Z}_\infty|\mathcal{A}_n) \geq \bar{Z}_n$ almost surely, $\mathbb{E}(\bar{W}_\infty|\mathcal{A}_n) \leq \bar{W}_n$ almost surely.

The continuity and monotonicity of $V(\cdot, Y)$ yields that $V(Z_n, Y) \uparrow V(m(X;\theta^*), Y)$ and $V(W_n, Y) \downarrow V(m(X;\theta^*), Y)$ almost surely as $n \to \infty$. Finally, the dominated convergence theorem yields, as $n \to \infty$,

$$\begin{aligned}0 \leq \mathbb{E}(\bar{W}_\infty - \bar{Z}_\infty) &= \mathbb{E}(\mathbb{E}(\bar{W}_\infty - \bar{Z}_\infty|\mathcal{A}_n)) \\ &\leq \mathbb{E}(\bar{W}_n - \bar{Z}_n) = \mathbb{E}(V(W_n, Y) - V(Z_n, Y)) \to 0.\end{aligned}$$

Therefore, $\bar{W}_\infty = \bar{Z}_\infty = 0$ almost surely. In particular, $\bar{W}_n - \bar{Z}_n \to 0$ almost surely.

Monotonicity of $V(Y, \cdot)$ implies that

$$\bar{Z}_n \leq \mathbb{E}(V(m(X;\theta^*), Y)|\mathcal{A}_n) \leq \bar{W}_n \text{ almost surely.}$$

Finally, Lévy's Zero-One-Law (Lévy, 1954) yields that

$$\mathbb{E}(V(m(X;\theta^*), Y)|\mathcal{A}_n) \to \mathbb{E}(V(m(X;\theta^*), Y)|\mathcal{A}_\infty)$$

almost surely as $n \to \infty$, where $\mathcal{A}_\infty = \bigcup_n \mathcal{A}_n = \sigma(m(X;\theta^*))$. □

Remark 4.A.2. Theorem 4.3.3 continues to hold for other functionals defined via an identification function. However, some additional assumptions are required. We need to assume that the identification function V is increasing in its first argument, that $V(m(X, \theta^*, Y))$ is integrable, and that $V(\cdot, Y)$ is continuous in the point $m(X;\theta^*)$ almost surely. Moreover, since $\mathbb{E}V(Z_0, Y)$ and $\mathbb{E}V(W_0, Y)$ can not necessarily be bounded from

above by $\mathbb{E}V(m(X,\theta^*),Y)$ for other identification functions, we need the additional assumption that $\mathbb{E}V(Z_0,Y)$ and $\mathbb{E}V(W_0,Y)$ are integrable.

Proof of Proposition 4.4.1. Note that

$$\arg\min_\theta \mathbb{E}S_\eta(m(X;\theta),Y) = \arg\min_\theta \mathbb{E}\mathbb{1}\{\eta \le m(X;\theta)\}(\eta - g(X)).$$

Our assumptions ensure that the expected elementary score is minimized if, and only if, $\{x : m(x;\theta) \ge \eta\} = \{x : g(x) \ge \eta\}$.

(a) Let θ be Pareto-optimal and assume θ is not a minimizer of any elementary loss. Thus, for every η, we have

$$\{x : m(x:\theta) \ge \eta\} \ne \{x : g(x) \ge \eta\}.$$

Let x, x' be such that

$$\begin{aligned} m(x;\theta) \ge \eta \quad &\text{but} \quad g(x) < \eta, \\ m(x';\theta) < \eta \quad &\text{but} \quad g(x') \ge \eta. \end{aligned}$$

Since $m(\cdot,\theta)$ is increasing, it follows that $x' < x$. But also $g(\cdot)$ is increasing which implies $x < x'$ contradicting the previous conclusion. Therefore, for all η, we have either

$$\begin{aligned} &\{x : m(x;\theta) \ge \eta\} \subseteq \{x : g(x) \ge \eta\}, \\ &\text{or } \{x : m(x;\theta) \ge \eta\} \supseteq \{x : g(x) \ge \eta\}. \end{aligned}$$

Suppose that there exist $\eta_1 \le \eta_2$ such that

$$\begin{aligned} &\{x : m(x;\theta) \ge \eta_1\} \subsetneq \{x : g(x) \ge \eta_1\}, \\ &\text{but } \{x : m(x;\theta) \ge \eta_2\} \supsetneq \{x : g(x) \ge \eta_2\}. \end{aligned}$$

Then, there exist x_1, x_2 with

$$\begin{aligned} m(x_1;\theta) < \eta_1 \quad &\text{and} \quad g(x_1) \ge \eta_1, \\ m(x_2;\theta) \ge \eta_2 \quad &\text{and} \quad g(x_2) < \eta_2. \end{aligned}$$

Since $m(\cdot;\theta)$ is increasing, we have $x_1 < x_2$. Therefore,

$$m(x_1;\theta) < \eta_1 \le g(x_1) \le g(x_2) < \eta_2 \le m(x_2;\theta).$$

This implies that $g(\cdot) - m(\cdot, \theta)$ changes its sign from positive to negative in $[x_1, x_2]$. Define

$$x^* := \inf\{x \in [x_1, x_2] : m(x; \theta) - g(x) > 0\}.$$

Then, $\eta^* := \lim_{x \downarrow x^*} g(x)$ satisfies

$$\{x : m(x; \theta) \geq \eta^*\} = \{x : g(x) \geq \eta^*\}.$$

This contradicts what we have previously shown. Moreover, note that we have only used that $m(\cdot, \theta)$ and $g(\cdot)$ are increasing. Therefore, we can swap the roles of $m(\cdot, \theta)$ and $g(\cdot)$ to obtain that either

$$\{x : m(x; \theta) \geq \eta\} \subsetneq \{x : g(x) \geq \eta\}, \quad \text{for all } \eta \in \mathbb{R},$$

or

$$\{x : m(x; \theta) \geq \eta\} \supsetneq \{x : g(x) \geq \eta\}, \quad \text{for all } \eta \in \mathbb{R}.$$

First, assume that $\{x : m(x; \theta) \geq \eta\} \subsetneq \{x : g(x) \geq \eta\}$ for all $\eta \in \mathbb{R}$. This implies that

$$\frac{1}{\theta_1}(\eta - \theta_0) > g^{-1}(\eta), \quad \text{for all } \eta.$$

Since $g^{-1}(\cdot)$ is increasing, our assumptions ensure that

$$\inf\left\{\frac{1}{\theta_1}(\eta - \theta_0) - g^{-1}(\eta) : \eta \in \mathbb{R}\right\} > 0.$$

Thus, we can choose $\delta > 0$ such that $\bar{\theta} = (\theta_0 + \delta, \theta_1)$ satisfies

$$\{x : m(x; \theta) \geq \eta\} \subsetneq \{x : m(x; \bar{\theta}) \geq \eta\} \subseteq \{x : g(x) \geq \eta\}, \quad \text{for all } \eta \in \mathbb{R},$$

and such that $\{x : m(x; \bar{\theta}) \geq \eta\} = \{x : g(x) \geq \eta\}$ for at least one η. Therefore,

$$\begin{aligned}
&\mathbb{E}S_\eta(m(X; \theta), Y) - \mathbb{E}S_\eta(m(X; \bar{\theta}), Y) \\
&= \mathbb{E}\Big(\Big(\mathbb{1}\{m(X; \theta) \geq \eta\} - \mathbb{1}\{m(X; \bar{\theta}) \geq \eta\}\Big)\Big(\eta - g(X)\Big)\Big) \geq 0,
\end{aligned}$$

because $\Big(\mathbb{1}\{m(X; \theta) \geq \eta\} - \mathbb{1}\{m(X; \bar{\theta}) \geq \eta\}\Big) = 0$ whenever $(\eta - g(X)) < 0$. The inequality is strict for at least one η, so that θ is strictly dominated by $\bar{\theta}$. Now, assume that $\{x : m(x; \theta) \geq \eta\} \supsetneq \{x : g(x) \geq \eta\}$ for all $\eta \in \mathbb{R}$. Thus, we have

$$\frac{1}{\theta_1}(\eta - \theta_0) < g^{-1}(\eta), \quad \text{for all } \eta.$$

Again our assumptions ensure

$$\sup\left\{\frac{1}{\theta_1}(\eta-\theta_0)-g^{-1}(\eta):\eta\in\mathbb{R}\right\}<0.$$

Therefore, we can choose $\delta>0$ such that $\bar{\theta}=(\theta_0-\delta,\theta_1)$ satisfies

$$\{x:m(x;\theta)\geq\eta\}\supsetneq\{x:m(x;\bar{\theta})\geq\eta\}\supseteq\{x:g(x)\geq\eta\},\quad\text{for all }\eta\in\mathbb{R},$$

and such that $\{x:m(x;\bar{\theta})\geq\eta\}=\{x:g(x)\geq\eta\}$ for at least one η. Following the same line of argumentation as in the first case, we can conclude that θ is strictly dominated by $\bar{\theta}$.

(b) Let θ be such that $\liminf_{\eta\to\infty}\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta)=0$. Observe that

$$0=\liminf_{\eta\to\infty}\theta_1^{-1}(\eta-\theta_0)-g^{-1}(\eta)=\liminf_{\eta\to\infty}\left(1-\frac{\theta_0}{\eta}-\frac{g^{-1}(\eta)}{\eta}\theta_1\right)\eta$$

implies that $\liminf_{\eta\to\infty}(1-\theta_1 g^{-1}(\eta)/\eta)=0$ or equivalently

$$\limsup_{\eta\to\infty}\frac{g^{-1}(\eta)}{\eta}=\limsup_{\eta\to\infty}\frac{1}{g'(g^{-1}(\eta))}=\frac{1}{\liminf_{\eta\to\infty}g'(g^{-1}(\eta))}=\frac{1}{\theta_1}.$$

Therefore, $\theta_1=\liminf_{\eta\to\infty}g'(g^{-1}(\eta))$ and

$$\liminf_{n\to\infty}(\eta-\theta_1 g^{-1}(\eta))=\theta_0$$

are uniquely determined. The statement for the other limits follows follows with a similar argument. □

Proof of Theorem 4.4.2. (a) If follows from Proposition 4.4.1 (a), that we only need to consider minimizers of some elementary loss. We first show that the parameters given by (4.1) are indeed Pareto-optimal. For $\eta=g(x^0)$, we have

$$\begin{aligned}\mathbb{E}S_\eta(m(X;\theta),Y)=&\mathbb{E}\Big(\mathbb{1}\{g(x^0)\leq\theta_0+\theta_1X\}\Big(g(x^0)-g(X)\Big)\Big)\\&-\mathbb{E}\Big(\mathbb{1}\{g(x^0)\leq Y\}\Big(g(x^0)-Y\Big)\Big)\end{aligned}$$

For the specific choice of θ given by (4.1), we have

$$g(x^0)\leq\theta_0+\theta_1X,\quad\text{almost surely,}$$

if, and only if,

$$0\leq g'(x^0)(X-x^0),\quad\text{almost surely.}$$

By assumption $g' > 0$, so that we conclude

$$\begin{aligned} g(x^0) \leq \theta_0 + \theta_1 X, \quad \text{almost surely,} \quad &\iff \quad x^0 \leq X, \quad \text{almost surely,} \\ &\iff \quad g(x^0) \leq g(X), \quad \text{almost surely.} \end{aligned}$$

Hence, θ given by (4.1) is indeed the best choice for $\eta = g(x^0)$ which implies that θ given by (4.1) is a minimizer for $\eta = g(x^0)$. It remains to argue, that θ is indeed Pareto-optimal. To this end, notice that $m(\cdot;\theta)$ with θ given by (4.1) is the tangent to g at x^0. If there exists $x^1 \neq x^0$ with $m(x^1;\theta) = g(x^1)$, then θ minimizes $\mathbb{E}S_\eta(\cdot, Y)$ for $\eta \in \{g(x^0), g(x^1)\}$ and is uniquely determined by the system of equations

$$\begin{aligned} g(x^0) &= \theta_0 + \theta_1 x^0 \\ g(x^1) &= \theta_0 + \theta_1 x^1. \end{aligned}$$

Thus, there cannot exist $\theta' \neq \theta$ with $m(X;\theta')$ also minimizing $\mathbb{E}S_\eta(\cdot, Y)$ for $\eta \in \{g(x^0), g(x^1)\}$. Therefore, θ is Pareto-optimal. If no such x^1 exists, we have

$$\begin{aligned} m(x;\theta) > g(x), \quad &\text{for all } x \neq x^0, \text{ or} \\ m(x;\theta) < g(x), \quad &\text{for all } x \neq x^0. \end{aligned}$$

Assume the first case occurs, the only candidates θ' to strictly dominate θ are those with $m(x^0;\theta') = g(x^0)$ and $m(x^1;\theta') = g(x^1)$ for some $x^1 \neq x^0$. Assume $x_1 > x_0$, the other case is analogous. Because $m(x;\theta)$ is a tangent to g in x^0, for $\eta < g(x^0)$, we have that

$$\{x : m(x;\theta') \geq \eta\} \supseteq \{x : m(x;\theta) \geq \eta\} \supsetneq \{x : g(x) \geq \eta\}.$$

The first inclusion has to be strict for at least one η, so that for this specific η we have

$$\begin{aligned} &\mathbb{E}S_\eta(m(X;\theta'), Y) - \mathbb{E}S_\eta(m(X;\theta), Y) \\ &= \mathbb{E}\Big(\Big(\mathbb{1}\{m(X;\theta') \geq \eta\} - \mathbb{1}\{m(X;\theta) \geq \eta\}\Big)\Big(\eta - g(X)\Big)\Big) > 0, \end{aligned}$$

Hence, $\mathbb{E}S_\eta(m(x;\theta'), Y) > \mathbb{E}S_\eta(m(x;\theta), Y)$ so that θ is Pareto-optimal.

We now show that the parameters given by (4.2) are indeed Pareto-optimal. Observe that $m(X;\theta)$ with θ given by (4.2) minimizes $\mathbb{E}S_\eta(\cdot, Y)$ for $\eta \in \{g(x^1), g(x^2)\}$. Moreover, our assumptions ensure that the system of equations (4.2) has a unique solution, so that the parameter θ is uniquely determined. Therefore, there can not exist $\theta' \neq \theta$ with $m(X;\theta')$ also minimizing $\mathbb{E}S_\eta(\cdot, Y)$ for $\eta \in \{g(x^1), g(x^2)\}$ which implies that θ is Pareto-optimal.

Finally, we show that parameters that minimize some elementary loss for exactly one η and are not of the form (4.1) are not Pareto optimal. If $m(x;\theta')$ realizes only

one superlevel set of g, then $g(\cdot) - m(\cdot;\theta')$ changes sign exactly once. Let η' be such that $\{x : m(x;\theta') \geq \eta'\} = \{x : g(x) \geq \eta'\}$. Then, either

$$\begin{aligned} \{x : m(x;\theta') \geq \eta'\} &\supsetneq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' > \eta, \\ \{x : m(x;\theta') \geq \eta'\} &\subsetneq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' < \eta, \end{aligned}$$

or

$$\begin{aligned} \{x : m(x;\theta') \geq \eta'\} &\subsetneq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' > \eta, \\ \{x : m(x;\theta') \geq \eta'\} &\supsetneq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' < \eta. \end{aligned}$$

Assume that the first case occurs. Then, because g is increasing, we can slightly rotate the line $m(x;\theta')$ so that the rotated version $m(x;\theta)$ satisfies $m(x';\theta) = g(x')$ for $x' = \arg\min_x |g(x) - m(x;\theta')|$. The resulting model $m(x;\theta)$ assumes an additional superlevel set of g. Then

$$\begin{aligned} \{x : m(x;\theta') \geq \eta'\} &\supsetneq \{x : m(x;\theta) \geq \eta'\} \supseteq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' > \eta, \\ \{x : m(x;\theta') \geq \eta'\} &\subsetneq \{x : m(x;\theta) \geq \eta'\} \subseteq \{x : g(x) \geq \eta'\}, \quad \text{for } \eta' < \eta \end{aligned}$$

so that $\mathbb{E}S_\eta(m(x;\theta'), Y) - \mathbb{E}S_\eta(m(x;\theta), Y) \geq 0$, where the inequality is strict for $\eta = g(x')$. The argument for the second case is similar.

(b) The claim follows from part (a) and from Proposition 4.4.1 (b). Indeed, if one of the limits equals zero, then the corresponding parameter θ is uniquely determined. Thus, there can be at most r additional Pareto-optimal parameters. □

Proof of Proposition 4.4.5. The optimal superlevel sets for an isotonic function on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ can be obtained from the pool-adjacent violators algorithm (PAVA); see Barlow et al. (1972), Jordan et al. (2019). The assumption that the model assumes two separate superlevel sets of $\hat{g}_{PAV}$ yields a system of two linear equations with a unique solution. Thus, this assumption characterizes the model parameter of $m(x;\theta)$ uniquely. Hence, there exists no other parameter being optimal for both of the above η-superlevel sets. Hence, θ is indeed Pareto-optimal.

The argument that these are the only Pareto-optimal parameters follows with the same argument as in the proof of Theorem 4.4.2. That only superlevel sets not equal to $\emptyset$ and $\{X_1, \ldots, X_n\}$ suffice for Pareto-optimality follows with a similar argument. □

Bibliography

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26:641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, London.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika*, 46(1/2):36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives. II. *Biometrika*, 46:328–335.

Bauer, H. (1974). *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. De Gruyter, Berlin.

Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780.

Brümmer, N. and Du Preez, J. (2013). The PAV algorithm optimizes binary proper scoring rules. *arXiv:1304.2331*.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26:607–616.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019). Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359.

Dawid, A. P. (2016). Contribution to the discussion of "Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings" by Ehm,

W., Gneiting, T., Jordan, A. and Krüger, F. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(3):505–562.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(3):505–562.

Feiler, M. J. and Ajdler, T. (2019). Model uncertainty in financial forecasting. *arXiv:1912.10813*.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Fissler, T. and Ziegel, J. F. (2019). Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics*, 13(1):1166–1211.

Franses, P. H. (1998). *Time Series Models for Business and Economic Forecasting*. Cambridge University Press.

Frongillo, R. and Kash, I. A. (2020). Elicitation Complexity of Statistical Properties. *Biometrika*. In press, https://doi.org/10.1093/biomet/asaa093.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, New York.

Grünwald, P. and Roos, T. (2019). Minimum description length revisited. *International Journal of Mathematics for Industry*, 11:1930001, 29.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594.

Gurney, A. J. T. and Griffin, T. G. (2011). Pathfinding through congruences. In *Relational and Algebraic Methods in Computer Science*, volume 6663, pages 180–195. Springer, Heidelberg.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2019). Isotonic distributional regression. *arXiv:1909.03725*.

Holland, M. J. (2019). Distribution-robust mean estimation via smoothed random perturbations. *arXiv:1906.10300*.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

Huggins, J. H. and Miller, J. W. (2019). Robust inference and model criticism using bagged posteriors. *arXiv:1912.07104*.

Jordan, A. I., Mühlemann, A., and Ziegel, J. F. (2019). Optimal solutions to the isotonic regression problem. *arXiv:1904.04761*.

Kyng, R., Rao, A., and Sachdeva, S. (2015). Fast, provable algorithms for isotonic regression in all L_p-norms. In *Advances in Neural Information Processing Systems 28*, pages 2719–2727. Curran Associates, Inc., Red Hook.

Lambert, N. S. (2019). Elicitation and evaluation of statistical forecasts. *Preprint*. Stanford University, Stanford. (Available from http://ai.stanford.edu/~nlambert/papers/elicitability.pdf).

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138.

Lévy, P. (1954). *Théorie de l'addition des variables aléatoires*. Collection des monographies des probabilités. Gauthier-Villars, Paris.

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2019). Focused Bayesian prediction. *arXiv:1912.12571*.

Luss, R. and Rosset, S. (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics*, 23(1):192–210.

Luss, R. and Rosset, S. (2017). Bounded isotonic regression. *Electronic Journal of Statistics*, 11(2):4488–4514.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, 19:724–740.

Miles, R. E. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 46:317–327.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1):24–49.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4):1901–1911.

Osband, K. H. (1985). *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38(4):796–809.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2):388–413.

Polonik, W. (1998). The silhouette, concentration functions and ML-density estimation under order restrictions. *The Annals of Statistics*, 26(5):1857–1877.

Robertson, T. and Wright, F. T. (1973). Multiple isotonic median regression. *The Annals of Statistics*, 1:422–432.

Robertson, T. and Wright, F. T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *The Annals of Statistics*, 8(3):645–651.

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks*, 11(6):1263–1271.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801.

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. (2014). Elicitation and identification of properties. In *Proceedings of the 27th Conference on Learning Theory*, pages 1–45.

Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica*, 71(2):450–470.

Thomas, O. and Corander, J. (2019). Diagnosing model misspecification and performing generalized Bayes' updates via probabilistic classifiers. *arXiv:1912.05810*.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, 20(3):360–380.

van Eeden, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions*. Mathematical Centre, Amsterdam.

Ziegel, J. F. (2016a). Coherence and elicitability. *Mathematical Finance*, 26(4):901–918.

Ziegel, J. F. (2016b). Contribution to the discussion of "Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings" by Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(3):505–562.

Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2020). Robust forecast evaluation of expected shortall. *Journal of Financial Econometrics*, 18:95–120.

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Mühlemann Anja

Registration Number: 11-108-982

Study program: PhD in Statistics

Bachelor ☐ Master ☐ Dissertation ☑

Title of the thesis: The Role of Loss Functions in Regression Problems

Supervisor: Prof. Dr. Johanna F. Ziegel

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of September 5th, 1996 and Article 69 of the University Statute of June 7th, 2011 is authorized to revoke the doctoral degree awarded on the basis of this thesis.
For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with theses submitted by others.

Place/Date Bern 1. März 2021

Signature A. Mühlemann